The Uses of Molecular Dating Analyses in Evolutionary Studies, with Examples from the Angiosperms

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Frank Kaspar Rutschmann

von

Zürich ZH

Promotionskomitee:

Prof. Dr. Elena Conti (Vorsitz)

Prof. Dr. Peter Linder

Dr. Torsten Eriksson

Zürich 2006

Acknowledgements

First of all, I would like to express my gratitude towards my supervisor, Elena Conti. The last four years would not have been as exciting and fruitful without your full support. I really appreciate your style of supervision which gave me the freedom to explore the things I was fascinated about. I am very impressed by your enthusiasm and your dedication to all sorts of biological questions, but especially by your social empathy. Your office door was always open for all kinds of questions - even beyond biology.

I would also like to thank Peter Linder - it was a great pleasure to collaborate with you, and your sense of humor made the struggle with statistics and manuscripts much more fun.

A great deal of laughter was also initiated by my fantastic and very patient office mates Merran Matthews, Josephine Jacksch and Gabriele Salvo, who often supplied me with mental and physical calories. But also Timo van der Niet, Sandro Wagen, Zsófia Hock and Péter Szövényi, Evelin Pfeifer, Urs Landergott, Brigitte Marazzi, Claudia Winteler, Chloé Galley, Rolf Rutishauser, Niklaus Müller, Alex Bernhard, Jakob Schneller, Burgi Liebst, Melanie Ranft, Rosemarie Sigrist, Reto Nyffeler, Daniel Heinzmann, Jürg Schönenberger, Maria von Balthazar, and many other members of the institute contributed to an enjoyable atmosphere and many cheerful discussions. The staff of the secretariat, the library, and the botanical garden made every journey outside the office a pleasure and provided an efficient working environment, especially Corinne Burlet, Claudio Brun, Barbara Seitz, Elena Benetti, Nadine Kofmehl, Elisabeth Schneeberger, and Daniel Schlagenhauf.

Finally, I would like to thank my family and friends, who supported me in many different ways. Thank you so much, Corinne Furrer, Anne and Christian Rutschmann, Dieter Rutschmann, Annette and Jochen Wild, Andri Schläpfer and Linda Zehnder, Renata and Hansjörg Schlaepfer, Philip Moline and Saskia Lampert, Nikolaus Amrhein, Annemarie and Niklaus Kohler-Pfister, Frances Athanasiou, Torsten Eriksson, Stephan Lange, Karen Hilzinger, Christoph Eichenberger, Beat Seiler, Lilo Michel, Lukas and Thomas Furrer, Dominique Sirena, Corsin Müller, Florian Steiner, and Jonas Wittwer.

Last but not least I owe much to my parents who have supported me in all conceivable ways. I can sincerely say that without you, this project would not have been possible. I therefore would like to dedicate this work to my mother, Margrit Rutschmann, and my late father, Peter Rutschmann.

Table of Contents

Publication list

The following publications stem from my years as a Ph.D. student at the Institute of Systematic Botany, University of Zurich, Switzerland (2002-2006). They represent ten articles, of which I have written five as a first or single author. Four papers are already published. The most recent work is listed first.

Rutschmann, F., J. Schönenberger, H.P. Linder, and E. Conti. Tempo and mode of diversification in the African Penaeaceae and related taxa. *In preparation.*

Barker N.P., P.H. Weston, and F. Rutschmann. Molecular dating of the radiation of the Proteaceae is only partially congruent with the breakup of Gondwana. *In preparation.*

Rutschmann F., T. Eriksson, K. Abu Salim, and E. Conti. Assessing calibration uncertainty in molecular dating: The assignment of fossils to alternative calibration points. *Submitted to Systematic Biology.*

Schlaepfer H. and F. Rutschmann. Deriving species responses from vegetation relevés: an iterative approach. *Submitted to Botanica Helvetica*.

Schönenberger, J., E. Conti, F. Rutschmann, and R. Dahlgren. Penaeaceae *in* K. Kubitzki, ed. The families and genera of vascular plants. Vol. 9. Springer, Berlin. *In press.*

Rutschmann F. 2006. Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Diversity and Distributions* **12**: 35-48.

Linder H.P., C. R. Hardy, and F. Rutschmann. 2005. Taxon sampling effects in molecular clock dating: An example from the African Restionaceae. *Molecular Phylogenetics and Evolution* **35**: 569–582.

Conti E., F. Rutschmann, T. Eriksson, K. Sytsma, and D. Baum. 2004. Calibration of molecular clocks and the biogeographic history of Crypteroniaceae: A reply to Robert G. Moyle. *Evolution* **58** (8): 1874-1876.

Rutschmann F. 2005. Bayesian molecular dating using PAML/multidivtime. A step-by-step manual, version 1.5 (July 2005). University of Zurich, Switzerland. Available at http://www.plant.ch.

Rutschmann F., T. Eriksson, J. Schönenberger, and E. Conti. 2004. Did Crypteroniaceae really disperse out-of-India? Molecular dating evidence from *rbc*L, *ndh*F, and *rpl*16 intron sequences. *International Journal of Plant Sciences* **165**(4 Suppl.): 69-83.

General introduction

In the quest of understanding evolutionary mechanisms and processes, accurate information about the timing of specific events in the past is needed. For example, it is important to know how fast retroviruses like the human immunodeficiency virus (HIV) change their genotype, in order to design new drugs or vaccines that will work also for future strains (Korber *et al.*, 2000). Or, to mention another fascinating example, it is of general interest - even beyond the life sciences - to know when the first humans started to diverge from the chimpanzees (between 6.3 and 5.4 million years ago; Patterson *et al.*, 2006).

Until forty years ago, information about the timing of past evolutionary events was only available in the form of fossils. Fossils are usually dated based on age estimates of the stratigraphic layer in which they are found, and they can be assigned to modern taxa by comparing the morphological characters preserved in the fossil with the traits of extant species (Hennig, 1969; Doyle and Donoghue, 1993). However, most organisms that once populated our planet have not been preserved in the fossil record, or their fossils have not been detected so far (Darwin, 1859). This implies that for most organisms, no direct information about their phylogenetic age is available.

In 1965, Zuckerkandl and Pauling postulated that the amount of difference between the DNA molecules of two species is a function of the time since their evolutionary separation. This was shown by comparing protein sequences (hemoglobins) from different species and further comparing amino acid substitution rates with ages estimated from fossils. Based on these early findings, standard techniques were developed to infer time information from molecular (DNA or protein) distances among species, giving origin, over time, to a range of different molecular dating methods (Magallón, 2004; Welch and Bromham, 2005; Penny, 2005).

In general, a molecular dating study begins with the reconstruction of the phylogenetic relationships among selected taxa. Then, specialized dating methods are used to transform the phylogenetic tree into an ultrametric tree. Whereas branch lengths in the phylogram represent the number of absolute molecular substitutions, they express relative time of lineage persistence in the ultrametric tree. The transformation of relative into absolute time requires some type of calibration, for which external time information – either stemming from fossils or from dated paleogeologic events - is assigned to at least one node of the phylogeny (Sanderson, 1998). The results of such a molecular dating analysis are estimates of divergence times for every lineage, summarized in a chronogram.

Under the assumption of a strict molecular clock, all branches of a phylogenetic tree evolve at the same, global substitution rate, which makes the second step, the reconstruction of an ultrametric tree, relatively easy (Felsenstein, 1981). However, it is known that variation in rates of nucleotide substitution, both along a lineage and between different lineages, is pervasive (Britten, 1986; Gillespie, 1986), especially if sequences of distantly related species are analyzed. Therefore, new "relaxed clock" molecular dating methods have been developed during the last ten years, which try to model rate variation among lineages (e.g., Penalized Likelihood, Sanderson 2002; Bayesian dating with multidivtime, Thorne and Kishino, 2002; relaxed clock model in BEAST, Drummond et al. 2006). This is a much more difficult task, because substitution rates and time have to be disentangled for every lineage. Furthermore, such modern dating methods also allow the incorporation of multiple calibration points (if available), and provide error estimates for the computed times and rates. In the first chapter of the present thesis (Rutschmann, 2006), I review the most common molecular dating methods by classifying them into three groups: i) methods that use a molecular clock and one global rate of substitution, ii) methods that correct for rate heterogeneity, and iii) methods that try to incorporate rate heterogeneity. Starting from a recent review by Magallón (2004), I expand to the most recent methods, emphasizing practical methodological details, providing useful comparison tables, and adding new links to the most important literature on molecular dating.

During the last fifteen years, molecular dating has become an important and increasingly popular tool for evolutionary biologists. For example, the timing of the eukaryotic evolution (Douzery *et al.*, 2004), the Early Cambrian origin of the main phyla of animals (Cambrian explosion; Wray *et al.*, 1996; Welch *et al.* 2005), and the replacement of dinosaurs by modern birds and mammals in the late Tertiary (Madsen, 2001) have all been investigated by estimating divergence times based on molecular sequence data. Also in plants, age estimates have been inferred at all taxonomic levels, for example for the plastid-containing eukaryotes (Yoon *et al.*, 2004), land plants (Sanderson, 2003), tracheophytes (Soltis *et al.*, 2002), angiosperms (Bell *et al.*, 2005; Magallón and Sanderson, 2005), the monocot-dicot divergence (Chaw *et al.*, 2004), Asterids (Bremer *et al.*, 2004), Dipsacales (Bell and Donoghue, 2005), Crypteroniaceae (Conti *et al.*, 2002), and *Fuchsia* (Berry *et al.*, 2004). However, in order to investigate evolutionary processes, age estimates exhibit their full potential only in combination with other historical evidence, such as the timing and sequence of paleotectonic events, climate reconstructions, or knowledge about the evolution of morphological and ecological traits.

For example, molecular dating estimates provide useful information for testing biogeographical hypotheses. In general, the dispersal of organisms can be explained either by long-distance-dispersal across existing barriers, or by disjunctions along ancient paleotectonic

corridors that subsequently disappeared (vicariance). Most probably, both dispersal and vicariance have played major roles in producing disjunct distribution patterns, and modern concepts like geodispersal (Liebermann, 2003; Ree *et al*., 2005) try to connect both models by focusing on the variable strength of the dispersal barrier over time. If the timeframe is known when two lineages with a disjunct distribution diverged, it is possible to search for compatible dispersal corridors formed by temporally reduced barriers which allowed a dispersal at that specific time. We search for such patterns in chapter two of this thesis by testing the out-of-India hypothesis connected to the dispersal of Crypteroniaceae, tropical rainforest trees currently occurring in Sri Lanka and Southeast Asia (Rutschmann *et al.*, 2004). In an earlier study, Conti *et al.* (2002) used molecular dating to test whether Crypteroniaceae were among other Gondwanan taxa that have been carried from Gondwana to Asia by rafting on the Indian plate. According to the hypothesis, Crypteroniaceae later dispersed "out-of-India" into South and Southeast Asia, after India collided with the Asian continent in the Early Tertiary. Conti *et al.* (2002) concluded in supporting this hypothesis for Crypteroniaceae, because both their phylogenetic reconstructions and their dating estimates of relevant nodes were concordant with the geologic history of the Indian Plate. However, because these conclusions were based on evidence from only one gene and limited taxon sampling, we repeated the analyses (see chapter two) by expanding both the taxon and gene sampling and comparing the results of three different molecular dating methods (clock-dependent Langley-Fitch, Langley and Fitch, 1974; Non Parametric Rate Smoothing, Sanderson, 1997; and Penalized Likelihood, Sanderson, 2002).

The out-of-India hypothesis is again addressed in the third thesis chapter (Conti *et al.*, 2004), which represents the answer to an opinion expressed by Robert G. Moyle in the International Journal of Organic Evolution (Moyle, 2004). Moyle reanalyzed the dataset from Conti *et al.* (2002) by using a different taxon sampling and another calibration point, criticizing the biogeographic interpretation proposed in Conti et al. (2002). In our reply, we highlight some weaknesses in Moyle's analytical procedure, at the same time offering some general reflections on the controversial issue of calibration in molecular dating analyses.

Time information from molecular dating can also be used to characterize diversification patterns stemming from periods of extraordinarily high or low speciation and/or extinction rates, such as rapid radiations. For example, the tempo and mode of a radiation can be linked to the simultaneous evolution of biotic and/or abiotic factors, such as key innovations or climatic and edaphic shifts. In chapter four of the present thesis, we investigate such patterns in a relatively small group of trees and shrubs which belong to the South and East African Penaeaceae, Oliniaceae, and Rhynchocalycaceae, closely related to the Crypteroniaceae studied in chapters two and three (Rutschmann *et al.*, *in prep*). Chronograms inferred in previous studies (Conti *et*

al., 2002; Rutschmann *et al.*, 2004) revealed long stem branches for both Penaeaceae and Oliniaceae, indicating a long time period between the origin and diversification of these lineages. The chronograms also suggested that this long phase was followed by a relatively short episode with rapid diversification, especially within Penaeaceae. Based on phylogenetic and molecular dating analyses, lineages through time plots, and inferences of ancestral states based on maximum likelihood, we aimed at reconstructing the tempo, mode, and possible reasons of diversification in Penaeaceae and Oliniaceae.

Despite their popularity, two main problems still plague the use of molecular dating methods. First, molecular dating relies entirely on the quality of calibration. A major source of uncertainty pertains to the assignment of fossils to specific nodes in a phylogeny, especially when alternative possibilities exist that can be equally justified on morphological grounds (Conti *et al.*, 2004; Magallón, 2004; Reisz and Müller, 2004). While widespread and challenging, this problem is often ignored in published papers. In the fifth chapter of this thesis (Rutschmann *et al.*, *submitted*), we used the fossil cross-validation procedure of Near and Sanderson (2004) in a novel way to assess uncertainty in fossil nodal assignment. More specifically, we used the phylogeny of Myrtales, six fossils, and 72 different combinations of calibration points to identify and characterize the assignments that produce the most internally consistent age estimates.

An additional major challenge in molecular dating is related to taxon sampling effects on the estimated ages. In the final chapter of this thesis (Linder *et al.*, 2005), we tested the sensitivity of three commonly used molecular dating methods to taxon sampling (Non Parametric Rate Smoothing, Sanderson, 1997; Penalized Likelihood, Sanderson, 2002; Bayesian dating with multidivtime, Thorne and Kishino, 2002). By using a nearly complete sample of the *Restio*-clade of the African grass-like Restionaceae (300 species, including 26 outgroup species), we formed nested subsets of 35, 51, 80, 120, and 150 species and performed molecular dating analyses based on the full dataset and all the subsets. We then compared the impact of the different dataset sizes and dating methods on the age estimates.

The present thesis comprises six chapters, four of them already published, that cover various aspects related to molecular dating, ranging from more methodological and experimental work to the application of different methods in the context of biological and evolutionary questions. I wish the reader an inspiring and pleasurable reading experience.

References

Bell, C. D., and M. J. Donoghue. 2005. Dating the dipsacales: Comparing models, genes, and evolutionary implications. *American Journal of Botany* **92**:284-296.

Bell, C. D., D. E. Soltis, and P. S. Soltis. 2005. The age of the Angiosperms: A molecular timescale without a clock. *Evolution* **59**:1245-1258.

Berry, P. E., W. J. Hahn, K. J. Sytsma, J. C. Hall, and A. Mast. 2004. Phylogenetic relationships and biogeography of *Fuchsia* (Onagraceae) based on noncoding nuclear and chloroplast DNA data. *American Journal of Botany* **94**:601-614.

Bremer, K., E. M. Friis, and B. Bremer. 2004. Molecular phylogenetic dating of asterid flowering plants shows early Cretaceous diversification. *Systematic Biology* **53**:496-505.

Britten, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**:1393-1398.

Chaw, S. M., C. C. Chang, H. L. Chen, and W.-H. Li. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *Journal of Molecular Evolution* **58**:424-441.

Conti, E., T. Eriksson, J. Schönenberger, K. J. Sytsma, and D. A. Baum. 2002. Early Tertiary out-of-India dispersal of Crypteroniaceae: evidence from phylogeny and molecular dating. *Evolution* **56**:1931-1942.

Conti, E., F. Rutschmann, T. Eriksson, K. J. Sytsma, and D. A. Baum. 2004. Calibration of molecular clocks and the biogeographic history of Crypteroniaceae: A reply to Moyle. *Evolution* **58**:1874-1876.

Darwin, C. 1859. On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life. John Murray, London.

Douzery, E. J. P., E. A. Snell, E. Bapteste, F. Delsuc, and H. Philippe. 2004. The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proceedings of the National Academy of Sciences USA* **101**:15386-15391.

Doyle, J. A., and M. J. Donoghue. 1993. Phylogenies and angiosperm diversification. *Paleobiology* **19**:141-167.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**:e88.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368-376.

Gillespie, J. H. 1986. Rates of molecular evolution. *Annual Review of Ecology, Evolution and Systematics* **17**:637-665.

Hennig, W. 1969. Die Stammesgeschichte der Insekten. Kramer, Frankfurt, Germany.

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**:1789-1796.

Langley, C. H., and W. Fitch. 1974. An estimation of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution* **3**:161-177.

Lieberman, B. S. 2003. Unifying theory and methodology in biogeography. *Evolutionary Biology* **33**

Linder, H. P., C. R. Hardy, and F. Rutschmann. 2005. Taxon sampling effects in molecular clock dating: an example from the African Restionaceae. *Molecular Phylogenetics and Evolution* **35**:569-582.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. Debry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. De Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**:610-614.

Magallón, S. A. 2004. Dating lineages: molecular and paleontological approaches to the temporal framework of clades. *International Journal of Plant Sciences* **165**:7-21.

Magallón, S. A., and M. J. Sanderson. 2005. Angiosperm divergence times: The effect of genes, codon positions, and time constraints. *Evolution* **59**:1653-1670.

Moyle, R. G. 2004. Calibration of molecular clocks and the biogeographic history of Crypteroniaceae. *Evolution* **58**:1871-1873.

Near, T. J., and M. J. Sanderson. 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **359**:1477-1483.

Patterson, N., D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature Advance online publication*: doi:10.1038/nature04789.

Penny, D. 2005. Relativity for molecular clocks. *Nature* **436**:183-184.

Ree, R. H., B. R. Moore, C. O. Webb, and M. J. Donoghue. 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution* **59**:2299-2311.

Reisz, R. R., and J. Müller. 2004. Molecular timescales and the fossil record: a paleontological perspective. *TRENDS in Genetics* **20**:237-241.

Rutschmann, F. 2006. Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Diversity and Distributions* **12**:35-48.

Rutschmann, F., T. Eriksson, K. Abu Salim, and E. Conti. Assessing calibration uncertainty in molecular dating: The assignment of fossils to alternative calibration points. *Submitted.*

Rutschmann, F., T. Eriksson, J. Schönenberger, and E. Conti. 2004. Did Crypteroniaceae really disperse out of India? Molecular dating evidence from *rbc*L, *ndh*F, and *rpl*16 sequences. *International Journal of Plant Sciences* **165**:69-83.

Rutschmann, F., J. Schönenberger, H. P. Linder, and E. Conti. Tempo and mode of diversification in the African Penaeaceae and related taxa. *In preparation.*

Sanderson, M. J. 1997. A nonparameteric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* **14**:1218-1231.

Sanderson, M. J. 1998. Estimating rate and time in molecular phylogenies: beyond the molecular clock? Pp. 242-264 *in* D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. Molecular systematics of plants II; DNA sequencing. Kluwer Academic Publishers, Norwell, MA.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a Penalized Likelihood approach. *Molecular Biology and Evolution* **19**:101-109.

Sanderson, M. J. 2003. Molecular data from 27 proteins do not support a Precambrian origin of land plants. *American Journal of Botany* **90**:954-956.

Soltis, P. S., D. E. Soltis, V. Savolainen, P. R. Crane, and T. G. Barraclough. 2002. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and evidence for molecular living fossils. *Proceedings of the National Academy of Sciences USA* **99**:4430-4435.

Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**:689-702.

Welch, J. J., and L. Bromham. 2005. Molecular dating when rates vary. *TRENDS in Ecology and Evolution* **20**:320-327.

Welch, J. J., E. Fontanillas, and L. Bromham. 2005. Molecular dates for the "Cambrian explosion": The influence of prior assumptions. *Systematic Biology* **54**:672-678.

Wray, G. A., J. S. Levinton, and L. S. Shapiro. 1996. Molecular evidence for deep precambrian divergences among metazoan phyla. *Science* **274**:568-573.

Yoon, H. S., J. D. Hackett, C. Ciniglia, G. Pinto, and D. Bhattacharya. 2004. A molecular timeline for the origin of photosynthetic Eukaryotes. *Molecular Biology and Evolution* **21**:809-818.

Zuckerkandl, E., and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. Pp. 97-166 *in* V. Bryson and H. Vogel, eds. Evolving Genes and Proteins. Academic Press, New York.

Chapter I.

Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times

Frank Rutschmann

Invited paper following the meeting "Recent Floristic Radiations in the Cape Flora" in Zurich, Switzerland, 3–5 July 2004, organized by Peter Linder, Chloé Galley, Cyril Guibert, Chris Hardy, Timo van der Niet, and Philip Moline, University of Zurich.

Published in

Diversity and Distributions 12: 35-48. 2006.

Abstract

This article reviews the most common methods used today for estimating divergence times and rates of molecular evolution. The methods are grouped into three main classes: a) methods that use a molecular clock and one global rate of substitution, b) methods that correct for rate heterogeneity, and c) methods that try to incorporate rate heterogeneity. Additionally, links to the most important literature on molecular dating are given, including articles comparing the performance of different methods, papers that investigate problems related to taxon, gene, and partition sampling, and literature discussing highly debated issues like calibration strategies and uncertainties, dating precision and the calculation of error estimates.

Key words: Divergence time estimation, molecular dating methods, rate heterogeneity, review.

Introduction

The use of DNA sequences to estimate the timing of evolutionary events is increasingly popular. The idea of dating evolutionary divergences using calibrated sequence differences was first proposed in 1965 by Zuckerkandl and Pauling (1965). The authors postulated that the amount of difference between the DNA molecules of two species is a function of the time since their evolutionary separation. This was shown by comparing protein sequences (hemoglobins) from different species and further comparing amino-acid substitution rates with ages estimated from fossils. Based on this central idea, molecular dating has been used in countless studies as a method to investigate mechanisms and processes of evolution. For example, the timing of the eukaryotic evolution (Douzery *et al.*, 2004), the Early Cambrian origin of the main phyla of animals (Cambrian explosion; Wray *et al.*, 1996; Smith and Peterson, 2002; Aris-Brosou and Yang, 2003), the replacement of dinosaurs by modern birds and mammals in the late Tertiary (Madsen, 2001), and the age of the last common ancestor of the main pandemic strain of human immunodeficiency virus (HIV; Korber *et al.*, 2000) have all been investigated using molecular dating. Also in plants, there are numerous studies where molecular dating methods have been used to investigate the timeframe of evolutionary events, e.g. for testing biogeographical hypotheses or to investigate the causes of recent radiations (for a more complete review see Sanderson *et al.*, 2004). For example, dating techniques have been applied on taxa from very different taxonomic levels, e.g. to infer the age of plastid-containing eukaryotes (Yoon *et al.*, 2004), land plants (Sanderson, 2003a), tracheophytes (Soltis *et al.*, 2002), angiosperms (Bell *et al.*, 2005; Wikström *et al.*, 2001, 2003; Magallón, 2001; Sanderson and Doyle, 2001), monocot-dicot divergence (Chaw *et al.*, 2004), Asterids (Bremer *et al.*, 2004), Myrtales (Sytsma *et al.*, 2004), Crypteroniaceae (Conti *et al.*, 2002), and *Fuchsia* (Berry *et al.*, 2004).

The goal of this article is to give a short overview on the most commonly used molecular dating methods. To allow for easier comparisons, the different methods for estimating divergence times are also summarized in tables 1-3.

The ideal case scenario:
A molecular clock and one global rate of substitution

In the special case of a molecular clock, all branches of a phylogenetic tree evolve at the same, global substitution rate. The clock-like tree is ultrametric, which means that the total distance between the root and every tip is constant.

Method 1: Linear regression (Nei, 1987; Li and Graur, 1991; Hillis *et al.*, 1996; Sanderson, 1998)

In an ultrametric tree, nodal depths can be converted easily into divergence times, because the molecular distance between each member of a sister pair and their most recent common ancestor is one half of the distance between the two sequences. If the divergence time for at least one node is known (calibration point), the global rate of substitution can be estimated and, based on that, divergence times for all nodes can be calculated by linear regression of the molecular distances (Li and Graur, 1991; Sanderson, 1998).

In other words: the observed number of differences D between two given sequences is a function of the constant rate of substitution r [subst. * site^{-1} * million years^{-1}] and the time t [million years] elapsed since the lineage exists. If we have one calibration point (e.g. if we can assign a fossil or geological event to one specific node in the tree), we can calculate the global substitution rate as follows: $r = D / 2t$. In a second step, we can use the global rate r to calculate the divergence time between any other two sequences: $T = D / 2r$.

In those cases where we have more than one calibration point, we can plot all calibration nodes in an age-genetic distance diagram, build a (weighted) regression line, whose slope is a function of the global substitution rate, and then interpolate (or extrapolate) the divergence times for the unknown nodes. The scatter of data points around the regression line provides then a confidence interval around estimated ages.

Although it is possible to estimate the global substitution rate r over an entire phylogeny (see methods 3 and 4), pairwise sequence comparisons have provided the most widely used approach to molecular dating until approximately five years ago, probably because the calculations can be done easily with any statistical or spreadsheet software.

Method 2: Tree-based mean path length method (Bremer and Gustafsson, 1997; Britton *et al.*, 2002)

The *mean path length method* estimates rates and divergence times based on the mean path length (MPL) between a node and each of its terminals. By calculating the MPL between a calibration node and all its terminals, and dividing it by the known age of the node, the global substitution rate is obtained. To calculate the divergence time of a node, its MPL is divided by the global rate. Although the calculation is simple enough to be done by hand, Britton *et al.* (2002) provide a Pascal program, named *PATH*, for the analysis of larger trees.

Method 3: Tree-based maximum likelihood clock optimization (Langley and Fitch, 1974; Sanderson, 2003b)

The *Langley-Fitch method* uses maximum likelihood (ML) to optimize the global rate of substitutions, starting with a phylogeny for which branch lengths are known. Using the optimized, constant rate, branch lengths and divergence times are then estimated. Finally, the results plus the outcome of a chi squared test of rate constancy are reported. The method is implemented in *r8s* (Sanderson, 2003b).

Method 4: Character-based maximum likelihood clock optimization (Felsenstein, 1981; Swofford *et al.*, 1996)

Under the assumption of rate constancy (and therefore under the constraints of ultrametricity), the global rate of substitution can also be optimized by ML directly from sequence data during the phylogenetic reconstruction. The likelihood is then a much more complex function of the data matrix, and the computing time is much higher than for the phylogenetic reconstruction without ultrametric constraints. Once the global rate of substitution is known, branch lengths and divergence times can be calculated.

The *ML clock optimization method* is implemented, at least partially, in *PAUP** (Swofford *et al.*, 2001), *DNAMLK* (part of the PHYLIP package); Felsenstein, 1993), *baseml* (part of the paml package; Yang, 1997), and other phylogenetic packages. It's perhaps the most widespread strategy commonly known as "enforcing the molecular clock". Depending on the software, the additional constrain of a fixed tree topology can be provided by the user, which reduces the complexity of the likelihood function and the computing time significantly.

The reality (in most cases):
Rate heterogeneity, or the relaxed clock

As sequences from multiple species began to accumulate during the 1970's, it became apparent that a clock is not always a good model for the process of molecular evolution (Langley and Fitch, 1974).Variation in rates of nucleotide substitution, both along a lineage and between different lineages, is known to be pervasive (Britten, 1986; Gillespie, 1986; Li, 1997). Several reasons are given for these deviations from the clock-like model of sequence evolution (some people call it relaxed or "sloppy" clock): a) generation time: a lineage with shorter generation time accelerates the clock because it shortens the time to accumulate and fix new mutations during genetic recombination (Ohta, 2002; but disputed by Whittle and Johnston, 2003); b) metabolic rate: organisms with higher metabolic rates have increased rates of DNA synthesis and higher rates of nucleotide mutations than species with lower metabolic rates (Gillooly *et al.* 2005; Martin and Palumbi, 1993); c) mutation rate: species-characteristic differences in the fidelity of the DNA replication or DNA repair machinery (Ota and Penny, 2003); d) effect of effective population size on the rate of fixation of mutations: the fixation of nearly-neutral alleles is expected to be the greatest in small populations (according to the nearly-neutral theory of DNA evolution; Ohta, 2002).

Because it is much easier to calculate divergence times under the clock model (with one global substitution rate), it is worth testing the data for clock-like behavior. This can be done by comparing how closely the ultrametric and additive trees fit the data. For example, the likelihood score of the best ultrametric tree can be compared with the (usually higher) likelihood score of the best additive tree and the difference between the two values (multiplied by two) can be checked for significance on a chi square table with $n-2$ degrees of freedom (where n is the number of terminals in the tree; Likelihood ratio test; Felsenstein, 1988, 1993, 2004; Muse and Weir, 1992). Other tests that can be applied to identify parts of the tree that show significant rate deviations are the relative rates test (Wu and Li, 1985) and the Tajima test (Tajima, 1993). However, all these clock tests lack power for shorter sequences and will detect only a relatively low proportion of cases of rate variation for the types of sequence that are typically used in molecular clock studies (Bromham *et al.*, 2000; Bromham and Penny, 2003). Failure to detect clock variation can cause systematic error in age estimates, because undetected rate variation can lead to significantly over- or underestimated divergence times (Bromham *et al.*, 2000). If the null hypothesis of a constant rate is rejected, or if we have evidence suggesting that test results should be treated with caution, we might conclude that rates vary across the tree; in such cases the use of

methods that try to model rate changes over the tree is necessary. This procedure is also supported by the fact that an increasing number of divergence time analyses show significant deviations from a molecular clock, especially if sequences of distantly related species are analyzed (e.g. different orders or families; Hasegawa *et al.*, 2003; Springer, 2003; Yoder and Yang, 2000). At least two groups of methods try to handle a relaxed clock: a) Methods that *correct* for rate heterogeneity before the dating; and b) methods that *incorporate* rate heterogeneity in the dating process, on the basis of specific rate change models.

1. Methods that *correct* for rate heterogeneity

The first set of methods described below correct for the observed rate heterogeneity by pruning branches or dividing the global rate into several rate classes (local rates). After this first step, which makes the trees (at least partially) ultrametric, they estimate rates and divergence times using the molecular clock as described above.

Method 5: Linearized trees (Li and Tanimura, 1987; Takezaki *et al.*, 1995; Hedges *et al.*, 1996)

The *linearized trees method* involves three steps: first, identify all branches in a phylogeny that depart significantly from rate constancy by using a statistical test (e.g. relative rates tests, Li and Tanimura, 1987; or two-cluster and branch-length tests, Takezaki et al, 1995). Then, selectively eliminate (prune) those branches. Finally, construct a tree with the remaining branches (the linearized tree) under the assumption of rate constancy. This procedure relies on eliminating data that do not fit the expected global rate behavior, and in many cases, this approach would lead to a massive elimination of data. Cutler (2000), describing the procedure as "taxon shopping", stated: "If one believes that rate overdispersion is intrinsic to the process of evolution (Gillespie, 1991) (...), then restricting one's analysis to taxa which happen to pass relative-rate tests is inappropriate".

Method 6: Local rates methods (Hasegawa *et al.*, 1989; Uyenoyama, 1995; Rambaut and Bromham, 1998; Yoder and Yang, 2000)

Apply two or more local molecular clocks on the tree by using a common model that characterizes the rate constancy on each part of the tree. One substantial difficulty is to identify correctly the branches or regions of a tree in which substitution rates significantly differ from the others; this difficulty explains why several methods of the "local rates type" exist. Usually, the

use of biological (e.g. similar life form, generation time, metabolic rate) or functional (e.g. gene function) information is used for their recognition. Also conspicuous patterns in the transition/transversion rate of some branches (Hasegawa *et al.*, 1989), the known differential function of alleles (Uyenoyama, 1995), the branch lengths obtained by ML under the absence of a molecular clock (Yoder and Yang, 2000; Yang and Yoder, 2003), or the rate constancy within a quartet of two pairs of sister groups (Rambaut and Bromham, 1998) can be used for the definition of local clock regions.

Probably the best known example of a local rates method is the *ML based local molecular clock approach (Hasegawa et al., 1989; Yoder and Yang, 2000; Yang and Yoder, 2003)*. This method pre-assigns evolutionary rates to some lineages while all the other branches evolve at the same rate. The local molecular clock model therefore lies between the two extremes of a global clock (assuming the same rate for all lineages) and the models that assume one independent rate for each branch (described below). The method allows for the definition of rate categories before the dating, which is a crucial and sensitive step for this method. Two different strategies can be used to pre-assign independent rates: a) definition of *rate categories*: the user pre-assigns rate categories to specific branches based on the branch length estimates obtained without the clock assumption. For example, three different rate categories are defined, one for the outgroup lineage with long branch lengths, another for a crown group with short branch lengths, and a third for all other branches; b) definition of *rank categories*: divide the taxa into several rate groups according to taxonomic ranks, e.g. order, suborder, family, and genus, based on the assumption that closely related evolutionary lineages tend to evolve at similar rates (Kishino *et al.*, 2001; Thorne and Kishino, 2002). After the definition of rate categories, the divergence times and rates for the different branch groups are estimated by ML optimization. The local molecular clock model is implemented in *baseml* (part of the *paml* package; Yang, 1997) and the program provides standard errors for estimated divergence times. *Baseml* does not (yet) allow for the specification of fossil calibrations as lower or upper limits on node ages, as in *r8s* or *multidivtime* (see below). So far, nodal constraints based on fossils have to be specified as fixed ages. The local molecular clock method implemented in *baseml* is now able to analyze multiple genes or data partitions with different evolutionary characteristics simultaneously and allows the branch group rates to vary freely among data partitions (since version 3.14). For example, the models allow some branches to evolve faster at codon position 1 while they evolve slower at codon position 2 (Yang and Yoder, 2003).

The quartet method (Rambaut and Bromham, 1998) implemented in *QDate* is one of the simplest local clock methods. The methods works with species quartets built by combining two pairs of species, each of which has a known date of divergence. For each pair, a rate can be

estimated, and this allows to estimate the date of the divergence between the pairs (age of the quartet). Because groups with undisputed relationships can be chosen, the methods avoids problems of topological uncertainties. On the other hand, it is difficult to combine estimates from multiple quartets in a meaningful way (Bromham *et al.*, 1998).

Another program that allows the user to assign different rates and substitution models to different parts of a tree is *Rhino* by Rambaut (2001). This ML local clock implementation has been used so far mainly for comparing substitution rates of different lineages by using likelihood ratio tests (e.g. Bromham and Woolfit, 2004).

A fourth implementation of the local molecular clock approach has been realized in the software *r8s* (Sanderson, 2003b). It follows the *Langley-Fitch method* described above (Method 3; Langley and Fitch, 1974), but instead of only using one constant rate of substitution, the method permits the user to specify multiple rate parameters and assign them to the appropriate branches or branch groups. After such a definition of rate categories, the divergence times and rates for the different branch groups are estimated by ML optimization.

A fifth program, *BEAST* (Drummond and Rambaut, 2003), uses Bayesian inference and the Markov chain Monte Carlo (MCMC) procedure to derive the posterior distribution of local rates and times. As the software does not require a starting tree topology, it is able to account for phylogenetic uncertainty. Additionally, it permits the definition of calibration distributions (such as normal, log-normal, exponential or gamma) instead of simple point estimates or age intervals.

2. Methods that estimate divergence times by *incorporating* rate heterogeneity

Methods that relax rate constancy must necessarily be guided by specifications about how rates are expected to change among lineages. Because rates and divergence times are confounded, it is not possible to estimate one without making assumptions regarding the other (Aris-Brosou and Yang, 2002). Recently, it has been questioned that divergence times without a molecular clock can be estimated consistently just by increasing the sequence lengths (Britton, 2005). However, available methods try to introduce rate heterogeneity on the basis of three different approaches: one is the concept of *temporal autocorrelation in rates* (see below 2a; Gillespie, 1991), another is the *stationary* process of rate change (see below 2b; Cutler, 2000), and a third is the *compound Poisson* process of rate change (see below 2c; Huelsenbeck *et al.*, 2000). All methods estimate branch lengths without assuming rate constancy, and then model the distribution of divergence times and rates by minimizing the discrepancies between branch

lengths and the rate changes over the branches. The methods differ not only in their models, but also in their strategy to incorporate age constraints (calibration points) into the analysis.

2a. Methods that model rate change according to the standard Poisson process and the concept of rate autocorrelation

An autocorrelation limits the speed with which a rate can change from an ancestral lineage to a descendant lineage (Sanderson, 1997). As the rate of substitution evolves along lineages, daughter lineages might inherit their initial rate from their parental lineage and evolve new rates independently (Gillespie, 1991). Temporal autocorrelation is an explicit *a priori* criterion to guide inference of among-lineage rate change and is implemented in several methods (Magallón, 2004). Readers who want to learn more about the theory of temporal autocorrelation are referred to publications by Takahata (1987), Gillespie (1991), Sanderson (1997), and Thorne *et al.* (1998).

Method 7: Nonparametric rate smoothing (Sanderson, 1997, 2003)

By analogy to smoothing methods in regression analysis, the *nonparametric rate smoothing (NPRS) method* attempts to simultaneously estimate unknown divergence times and smooth the rapidity of rate change along lineages (Sanderson, 1997, 2003). To smooth rate changes, the method contains a non-parametric function that penalizes rates that change too quickly from branch to neighboring branch, which reflects the idea of autocorrelation of rates. In other words: the local transformations in rate are smoothed as the rate itself changes over the tree by minimizing the ancestral-descendant changes of rate. Since the penalty function includes unknown times, an optimality criterion based on this penalty (the sum of squared differences in local rate estimates compared from branch to neighboring branch; least-squares method) permits an estimation of the divergence times. NPRS is implemented in *r8s* (Sanderson, 2003b) and *TreeEdit* (Rambaut and Charleston, 2002). With *r8s*, but not with TreeEdit, the user is able to add one or more calibration constraints to permit scaling of rates and times to absolute temporal units. A serious limitation of NPRS is that it tends to overfit the data, leading to rapid fluctuations in rate in regions of a tree that have short branches (Sanderson, 2003b). In *r8s*, but not in TreeEdit, two strategies to provide confidence intervals on the estimated parameters are available: a) a built-in procedure that uses the curvature of the likelihood surface around the parameter estimate (after Cutler, 2000); and b) the calculation of an age distribution based on chronograms generated from a large number of bootstrapped datasets. The central 95% of the age distribution provide the confidence interval (Efron and Tibshirani, 1993; Baldwin and Sanderson, 1998; Sanderson,

2003b). This robust, but time consuming procedure can be facilitated by using a collection of Perl scripts written by Eriksson (2002) called the *r8s-bootstrap-kit*.

Method 8: Penalized Likelihood (Sanderson, 2002, 2003)

Penalized likelihood (PL) combines likelihood and the nonparametric penalty function used in NPRS. This semi-parametric technique attempts to combine the statistical power of parametric methods with the robustness of non-parametric methods. In effect, it permits the specification of the relative contribution of the rate smoothing and the data-fitting parts of the estimation procedure: a roughness penalty can be assigned as smoothing parameter in the input file. The smoothing parameter can be estimated by running a cross-validation procedure, which is a data-driven method for finding the optimal level of smoothing. If the smoothing parameter is large, the function is dominated by the roughness penalty, and this leads to a clock-like model. If it is low, the smoothing will be effectively unconstrained (similar to NPRS). So far, PL is only implemented in *r8s* (Sanderson, 2003b). As with NPRS in *r8s*, the user is able to add one or more calibration constraints to permit scaling of rates and times to real units. The same two strategies for providing confidence intervals on the estimated parameters as for the NPRS method are also available for *PL*.

Method 9: Heuristic rate smoothing (AHRS) for ML estimation of divergence times (Yang, 2004)

The *heuristic rate-smoothing (AHRS)* algorithm for ML estimation of divergence times (Yang, 2004) has a number of similarities with *PL* and the two Bayesian dating methods described above. It involves three steps: a) estimation of branch lengths in the absence of a molecular clock; b) heuristic rate smoothing to estimate substitution rates for branches together with divergence times, and classification of branches into several rate classes; and c) ML estimation of divergence times and rates of the different branch groups. The *AHRS* algorithm differs slightly from PL: where Sanderson (2002) uses a Poisson approximation to fit the branch lengths, the *AHRS* algorithm uses a normal approximation of the ML estimates of branch lengths. Furthermore, the rate-smoothing algorithm in *AHRS* is used only to partition branches on each gene tree into different rate groups, and plays therefore a less significant role in this method than in *PL*. In contrast to the Bayesian approaches described above, *AHRS* optimizes rates, together with divergence times, rather than averaging over them in an MCMC procedure. Another difference to the Bayesian dating methods is that the *AHRS* algorithm does not need any prior for divergence times, which can be an advantage. On the other hand, it is not possible to specify

fossil calibrations as lower or upper bounds on node ages, as in *r8s* or *multidivtime* - so far, nodal constraints based on fossils have to be specified as fixed ages. The *AHRS* algorithm is implemented in the *baseml* and *codeml* programs, which are parts of the *paml* package (since version 3.14 final; Yang, 1997). Those programs provide standard errors for estimated divergence times. As *multidivtime,* the *AHRS* algorithm implemented in *baseml* is able to analyze multiple genes/loci with different evolutionary characteristics simultaneously.

Method 10: Bayesian implementation of rate autocorrelation in multidivtime (Thorne *et al.*, 1998; Kishino *et al.*, 2001; Thorne and Kishino, 2002)

The *Bayesian dating method* implemented in *multidivtime* (Thorne *et al.*, 1998; Kishino *et al.*, 2001; Thorne and Kishino, 2002) uses a fully probabilistic and high parametric model to describe the change in evolutionary rate over time and uses the Markov chain Monte Carlo (MCMC) procedure to derive the posterior distribution of rates and times. In effect, the variation of rates is addressed by letting the MCMC algorithm assign rates to different parts of the tree, and then sampling from the patterns that are possible. By this way, the MC techniques average over various patterns of rates along the tree. The result is a posterior distribution of rates and times derived from a prior distribution. For the assignments of rates to different branches in the tree, rates a drawn from a lognormal distribution, and a hyperparameter ν (also called Brownian motion constant) describes the amount of autocorrelation. The internal node age proportions are described as a dirichlet distribution, which represents the idea to model evolutionary lineages due to speciation, but is not intended as a detailed model of speciation and extinction processes. In practice, the most commonly used procedure is divided into three different steps and programs, and is described in more detail in a step-by-step manual (Rutschmann, 2005). 1) In the first step, model parameters for the F84 + G model (Kishino and Hasegawa, 1989; Felsenstein, 1993) are estimated by using the program *baseml*, which is part of the *PAML* package (Yang, 1997). 2) By using these parameters, the ML of the branch lengths is estimated, together with a variance-covariance matrix of the branch length estimates by using the program *estbranches* (Thorne *et al.*, 1998). 3) The third program, *multidivtime* (Kishino *et al.*, 2001; Thorne and Kishino, 2002), is then able to approximate the posterior distributions of substitution rates and divergence times by using a multivariate normal distribution of estimated branch lengths and running an MCMC procedure.

Multidivtime asks the user to specify several priors, such as the mean and the variance of the distributions for the initial substitution rate at the root node or the prospective age of the root node. Additionally, constraints on nodal ages can be specified as age intervals. The program

provides Bayesian credibility intervals for estimated divergence times and substitution rates. In contrast to *NPRS* and *PL* (implemented in *r8s*), *multidivtime* is able to account for multiple genes/loci with different evolutionary characteristics. Such a simultaneous analysis of multiple genes may improve the estimates of divergence times which are shared across genes.

Method 11: Phybayes (Aris-Brosou and Yang, 2001, 2002)

The *Phybayes* program (Aris-Brosou and Yang, 2001, 2002) is similar to the *multidivtime* Bayesian approach described above. It also uses a fully probabilistic and high parametric model to describe the change in evolutionary rate over time and uses the MCMC procedure to derive the posterior distribution of rates and times. But the method is more versatile in terms of possibilities of defining the prior distributions, as it allows for the usage of models that explicitly describe the processes of speciation and extinction. For the rates of evolution, it offers a choice of six different rate distributions to model the autocorrelated rate change from an ancestor to a descendent branch (lognormal, „stationarized"-lognormal, truncated normal, Ornstein-Uhlenbeck process, gamma, and exponential, plus the definition of two model-related hyperparameters), whereas in *multidivtime*, rates are always drawn from a lognormal distribution. The prior distribution of divergence times is generated by a process of cladogenesis, the generalized birth and death process with species sampling (Yang and Rannala, 1997), a model that assumes a constant speciation and extinction rate per lineage (*multidivtime* uses a dirichlet distribution of all internal node age proportions to generalize the rooted tree structure; Kishino *et al.*, 2001). In contrast to multidivtime, it's not possible to analyze multiple genes simultaneously with *Phybayes*, and the program does not allow for an *a priori* integration of nodal constraints (calibration points).

2b. Methods that model rate change with other concepts than rate autocorrelation

Method 12: Bayesian implementation of rate variation in BEAST (Drummond *et al.*, 2006)

Similar to *Phybayes*, the variable rate methods implemented in *BEAST* use Bayesian inference and the Markov chain Monte Carlo (MCMC) procedures to derive the posterior distribution of rates and times. In contrast, in addition to the autocorrelated models like the one implemented in *multidivtime*, a range of different, novel models have been implemented, where the rates are drawn from a distribution (with various distributions on offer; Drummond *et al.*, 2006). These models have a couple of interesting features: a) the parameters of the distributions can be estimated (instead of being specified), and b) the correlation of rates between adjacent branches can be tested (if > 0, this would indicate some inherence of rates). Another unique

feature of the software is that it does not require a starting tree topology, which allows it to account for phylogenetic uncertainty. Additionally, *BEAST* permits the definition of calibration distributions (such as normal, log-normal, exponential or gamma) to model calibration uncertainty instead of simple point estimates or age intervals. For the other, non-calibrated nodes, there is no specific process that describes the prior distribution of divergence times (they are uniform over a range from 0 to very large). *BEAST* allows the user to simultaneously analyze multiple datasets/partitions with different substitution models, and provides Bayesian credibility intervals.

Method 13: Overdispersed clock method (Cutler, 2000)

While all methods described so far assume fundamentally that the number of substitutions in a lineage is Poisson-distributed, the *overdispersed clock model* (Cutler, 2000) assumes that the number of substitutions in a lineage is stationary. Unlike a Poisson process, the variance in the number of substitutions will not necessarily be equal to the mean. Under this model, which treats rate changes according to a stationary process, ML estimates of divergence times can be calculated. The method is implemented in a c program, which is available directly from the author (Cutler, 2000). It is possible to incorporate multiple calibration points.

Method 14: Compound Poisson process method (Huelsenbeck *et al.*, 2000)

As all methods described above (with the exception of the *overdispersed clock model*; Cutler, 2000), the *compound Poisson process method* uses a model that assumes that nucleotide substitutions occur along branches of the tree according to a Poisson process. But in addition to the other models, it assumes that another, independent Poisson process generates events of substitution rate change. Therefore, this second Poisson process is superimposed on the primary Poisson process of molecular substitution (hence the name *compound Poisson* process), and introduces changes (in form of discrete jumps) in the rate of substitution in different branches of the phylogeny. Rates on the tree are then determined by the number of rate-change events, the point in the tree where they occur, and the magnitude of change at each event (Huelsenbeck *et al.*, 2000; Magallón, 2004). These parameters are estimated by using Bayesian inference (MCMC integration). One of the main advantages of treating rate variation as a *compound Poisson* process is that the model can introduce rate variation at any point of the phylogenetic tree; all other methods assume that substitution rates change only at speciation events (nodes; Huelsenbeck *et al.*, 2000). The method is implemented in a c program, but it's not meant for being available for the community so far as there is no documentation (yet).

Conclusions

Molecular dating is a rapidly developing field, and the methods generated so far are still far from being perfect (Sanderson *et al.*, 2004). Molecular dating estimates derived from different inference methods can be in conflict, and so can the results obtained with different taxon sampling, gene sampling, and calibration strategies (see below).

It should be clear that there is no single "best" molecular dating method; rather, all approaches have advantages and disadvantages. For example, the methods reviewed here differ in the type of input data they use and process: The first group of methods (*NPRS* and *r8s*) base their analysis on input phylograms with branch lengths. Therefore, they are not able to incorporate branch length errors or parameters of the substitution model into the dating analysis (this has to be done prior to the dating). On the other hand, these methods are fast and versatile, because they can process phylogenies generated from parsimony, likelihood or Bayesian analyses. The second group of methods (*multidivtime*, *Phybayes*, *AHRS*) use one "true" tree topology to assess rates and divergence times and estimate the branch lengths themselves. Therefore, they are able to account for the branch length errors described above, but still base their analyses on a fixed, user supplied tree topology. The third group of methods (*ML with clock* and *BEAST*) directly calculate ultrametric phylogenies based only on sequence data and model parameters, a procedure that also allows them to incorporate topological uncertainties. Computationally, this can be very expensive, especially in the case of a variable rate model, and with a high number of taxa.

As I have not tested all software packages and methods described here myself, this review is not a comparison based on practical experiments. However, the papers that first described these approaches always report on their performance on simulated and real data. Three papers that really compare some of the described methods on original data sets or simulated data have been published recently: Yang and Yoder (2003) compared methods 3, 5 and 8 (see Tables 1-3) by analyzing a Mouse Lemur dataset using multiple gene loci and calibration points, Pérez-Losada *et al.* (2004) compared methods 5, 7, 8, and 9 in their analysis of a nuclear ribosomal 18S dataset to test the evolutionary radiation of the Thoracian Barnacles by comparing different calibration points independently, and Ho *et al.* (2005b) compared the performance of methods 8, 10, and 11 by using simulated data. Currently, software developers are starting to integrate different methods in the same software (e.g. *baseml*, Yang, 1997; *BEAST*, Drummond *et al.*, 2006; and future versions of *MrBayes* > v3.1, Huelsenbeck and Ronquist, 2001). This recent trend is thus allowing users to try out different methods based on their own datasets, and then compare the results.

Although this review focuses on the dating methods themselves, at least a few links to key articles about more general or very specific issues related to molecular dating are given here: 1)

general reviews: Magallón (2004) wrote a comprehensive review of the theory of molecular dating, which also discusses paleontological dating methods and the uncertainties of the paleontological record. The classification of the molecular methods described here is based on her publication. Another recent review has been written by Sanderson *et al.* (2004), which discusses the advantages and disadvantages of Bayesian vs. smoothing molecular dating methods, summarizes the inferred ages of the major clades of plants, and lists many published dating applications that investigated recent plant radiations and/or tested biogeographical hypotheses. Finally, Welch and Bromham (2005), Bromham and Penny (2003), Arbogast *et al.* (2002), Wray (2001) wrote more general reviews on the issue of estimating divergence times. 2) For specific discussions about the crucial and controversial role of calibration, refer to the following papers: Where on the tree and how should we assign ages from fossils or geologic events?: Near and Sanderson (2004), Conti *et al.* (2004) and Rutschmann *et al.* (2004). How can we deal with the incompleteness of the fossil record?: Tavaré *et al.* (2002), Foote *et al.* (1999), and Foote and Sepkoski (1999). How should we constrain the age of the root and deal with the methodological handicap of asymmetric random variables in molecular dating?: Rodríguez-Trelles *et al.* (2002) and Sanderson and Doyle (2001). 3) For the recent debate about the precision of divergence time estimates, refer to the following papers: Should we extrapolate substitution rates accross different evolutionary timescales?: Ho *et al.* (2005a). How can we account for the various uncertainties related to the calibration and the dating procedure, how should we report and interpret error estimates, and should we use secondary calibration points?: Hedges and Kumar (2003), Graur and Martin (2004), Reisz and Müller (2004), and Hedges and Kumar (2004). 4) For questions related to the influence of taxon sampling on estimating divergence times under various dating methods, see Linder *et al.* (2005) and Sanderson and Doyle (2001). 5) For the influence of gene sampling read Heckman *et al.* (2001). 6) Theoretical problems and strategies connected to the molecular dating of supertrees are discussed in Vos and Mooers (2004) and Bryant *et al.* (2004). 6) For "special" dating problems like estimating the substitution rate when the ages of different terminals are known (e.g. from virus sequences that were isolated at different dates; implemented in the software *TipDate* and also in *BEAST*), refer to Rambaut (2000).

Finally, I would like to add a suggestion to those among us who write software: although the development of graphical user interfaces (GUI's) is certainly not a first priority, GUI's would simplify significantly molecular dating analyses for the average biologist. Modern tools (like the Open Source *Qt 4* C++ class library by *Trolltech*; http://www.trolltech.com) make the development of fast, native, and multiplatform GUI applications easier than ever before. I do not share the widespread concerns about stupid "analyse-by-click" users. On the contrary:

comprehensive user interfaces allow the user to explore and detect all the important features a method offers.

References

Arbogast, B. S., S. V. Edwards, J. Wakeley, P. Beerli, and J. B. Slwinski. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology, Evolution and Systematics* **33**:707-740.

Aris-Brosou, S. 2001. Phybayes: a program for phylogenetic analyses in a Bayesian framework. Department of Biology (Galton Laboratory), University College London, London, UK.

Aris-Brosou, S., and Z. Yang. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA Phylogeny. *Systematic Biology* **51**:703-714.

Aris-Brosou, S., and Z. Yang. 2003. Bayesian models of episodic evolution support a Late Precambrian explosive diversification of Metazoa. *Molecular Biology and Evolution* **20**:1947-1954.

Baldwin, B. G., and M. J. Sanderson. 1998. Age and rate of diversification of the hawaiian silversword alliance (Compositae). *Proceedings of the National Academy of Sciences USA* **95**:9402-9406.

Bell, C. D., D. E. Soltis, and P. S. Soltis. 2005. The age of the Angiosperms: A molecular timescale without a clock. *Evolution* **59**:1245-1258.

Berry, P. E., W. J. Hahn, K. J. Sytsma, J. C. Hall, and A. Mast. 2004. Phylogenetic relationships and biogeography of *Fuchsia* (Onagraceae) based on noncoding nuclear and chloroplast DNA data. *American Journal of Botany* **94**:601-614.

Bremer, K., E. M. Friis, and B. Bremer. 2004. Molecular phylogenetic dating of asterid flowering plants shows early Cretaceous diversification. *Systematic Biology* **53**:496-505.

Bremer, K., and M. H. G. Gustafsson. 1997. East Gondwana ancestry of the sunflower alliance of families. *Proceedings of the National Academy of Sciences USA* **94**:9188-9190.

Britten, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**:1393-1398.

Britton, T. 2005. Estimating divergence times in phylogenetic trees without a molecular clock. *Systematic Biology* **54**:500-507.

Britton, T., B. Oxelman, A. Vinnersten, and K. Bremer. 2002. Phylogenetic dating with conficence intervals using mean path lengths. *Molecular Phylogenetics and Evolution* **24**:58-65.

Bromham, L., and D. Penny. 2003. The modern molecular clock. *Nature Reviews Genetics* **4**:216-224.

Bromham, L., A. Rambaut, R. Fortey, A. Cooper, and D. Penny. 1998. Testing the Cambrian explosion hypothesis by using a molecular dating technique. *Proceedings of the National Academy of Sciences USA* **95**:12386-12389.

Bromham, L., and M. Woolfit. 2004. Explosive radiations and the reliability of molecular clocks: island endemic radiations as a test case. *Systematic Biology* **53**:758-766.

Bromham, L. D., A. Rambaut, M. D. Hendy, and D. Penny. 2000. The power of relative rates test depends on the data. *Journal of Molecular Evolution* **50**:296-301.

Bryant, D., C. Semple, and M. Steel. 2004. Supertree methods for ancestral divergence dates and other applications. Pp. 129-150 *in* O. R. P. Bininda-Emonds, ed. Phylogenetic supertrees: Combining information to reveal the tree of life. Kluwer Academic, Dordrecht, The Netherlands.

Chaw, S. M., C. C. Chang, H. L. Chen, and W.-H. Li. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *Journal of Molecular Evolution* **58**:424-441.

Conti, E., T. Eriksson, J. Schönenberger, K. J. Sytsma, and D. A. Baum. 2002. Early Tertiary out-of-India dispersal of Crypteroniaceae: evidence from phylogeny and molecular dating. *Evolution* **56**:1931-1942.

Conti, E., F. Rutschmann, T. Eriksson, K. J. Sytsma, and D. A. Baum. 2004. Calibration of molecular clocks and the biogeographic history of Crypteroniaceae: A reply to Moyle. *Evolution* **58**:1874-1876.

Cutler, D. J. 2000. Estimating divergence times in the presence of an overdispersed molecular clock. *Molecular Biology and Evolution* **17**:1647-1660.

Douzery, E. J. P., E. A. Snell, E. Bapteste, F. Delsuc, and H. Philippe. 2004. The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proceedings of the National Academy of Sciences USA* **101**:15386-15391.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**:e88.

Drummond, A. J., and A. Rambaut. 2003. BEAST v1.0. Available at http://evolve.zoo.ox.ac.uk/beast/.

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, New York.

Eriksson, T. 2002. The r8s-bootstrap-kit. Bergianska University, Stockholm, Sweden.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368-376.

Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22**:521-265.

Felsenstein, J. 1993. Phylogenetic Inference Package (PHYLIP), version 3.5. University of Washington, Seattle, WA.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Inc., Sunderland, MA.

Foote, M., J. P. Hunter, C. M. Janis, and J. J. Sepkoski Jr. 1999. Evolutionary and preservational constraints on origins of biologic groups: divergence times of Eutherian mammals. *Science* **283**:1310-1314.

Foote, M., and J. J. Sepkoski Jr. 1999. Absolute measures of the completeness of the fossil record. *Nature* **398**:415-417.

Gillespie, J. H. 1986. Rates of molecular evolution. *Annual Review of Ecology, Evolution and Systematics* **17**:637-665.

Gillespie, J. H. 1991. The causes of molecular evolution. Oxford University Press, Oxford, GB.

Gillooly, J. F., A. P. Allen, G. B. West, and J. H. Brown. 2005. The rate of DNA evolution: Effects of body size and temperature on the molecular clock. *Proceedings of the National Academy of Sciences USA* **102**:140-145.

Graur, D., and W. Martin. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *TRENDS in Genetics* **20**:80-86.

Hasegawa, M., H. Kishino, and T. Yano. 1989. Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. *Journal of Human Evolution* **18**:461-476.

Hasegawa, M., J. L. Thorne, and H. Kishino. 2003. Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes & Genetic Systems* **78**:267-283.

Heckman, D. S., D. M. Geiser, B. R. Eidell, R. L. Stauffer, N. L. Kardos, and S. B. Hedges. 2001. Molecular Evidence for the Early Colonization of Land by Fungi and Plants. *Science* **293**:1129-1133.

Hedges, S. B., and S. Kumar. 2003. Genomic clocks and evolutionary timescales. *TRENDS in Genetics* **19**:200-206.

Hedges, S. B., and S. Kumar. 2004. Precision of molecular time estimates. *TRENDS in Genetics* **20**:242-247.

Hedges, S. B., P. H. Parker, C. G. Sibley, and S. Kumar. 1996. Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**:226-229.

Hillis, D. M., B. K. Mable, and C. Moritz. 1996. Applications of molecular systematics: the state of the field an a look to the future. Pp. 515-543 *in* D. M. Hillis, C. Moritz and B. K. Mable, eds. Molecular Systematics. Sinauer, Sunderland, MA.

Ho, S. Y. W., M. J. Phillips, A. Cooper, and A. J. Drummond. 2005a. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular Biology and Evolution* **22**:1561-1568.

Ho, S. Y. W., M. J. Phillips, A. J. Drummond, and A. Cooper. 2005b. Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. *Molecular Biology and Evolution* **22**:1355-1363.

Huelsenbeck, J. P., B. Larget, and D. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* **154**:1879-1892.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* **17**:754.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution* **170-179**:170-179.

Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* **18**:352-361.

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**:1789-1796.

Langley, C. H., and W. Fitch. 1974. An estimation of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution* **3**:161-177.

Li, W.-H. 1997. Molecular Evolution. Sinauer Press, Sunderland MA, USA.

Li, W.-H., and D. Graur. 1991. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA.

Li, W.-H., and M. Tanimura. 1987. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326**:93-96.

Linder, H. P., C. R. Hardy, and F. Rutschmann. 2005. Taxon sampling effects in molecular clock dating: an example from the African Restionaceae. *Molecular Phylogenetics and Evolution* **35**:569-582.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. Debry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. De Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**:610-614.

Magallón, S. A. 2004. Dating lineages: molecular and paleontological approaches to the temporal framework of clades. *International Journal of Plant Sciences* **165**:7-21.

Magallón, S. A., and M. J. Sanderson. 2001. Absolute diversification rates in Angiosperm clades. *Evolution* **55**:1762-1780.

Martin, A. P., and S. R. Palumbi. 1993. Body size, metabolic rate, generation time and the molecular clock. *Proceedings of the National Academy of Sciences USA* **90**:4087-4091.

Muse, S. V., and B. S. Weir. 1992. Testing for equality of evoltuionary rates. *Genetics* **132**:269-276.

Near, T. J., and M. J. Sanderson. 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **359**:1477-1483.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Ohta, T. 2002. Near-neutrality in evolution of genes and in gene regulation. *Proceedings of the National Academy of Sciences USA* **99**:16134-16137.

Ota, R., and D. Penny. 2003. Estimating changes in mutational mechanisms of evolution. *Journal of Molecular Evolution* **57**:S233–S240.

Pérez-Losada, M., J. T. Høeg, and K. A. Crandall. 2004. Unraveling the evolutionary radiation of the Thoracian Barnacles using molecular and morphological evidence: a comparison of several divergence time estimation approaches. *Systematic Biology* **53**:244-264.

Rambaut, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**:395-399.

Rambaut, A. 2001. Rhino v1.1. Available at http://evolve.zoo.ox.ac.uk/.

Rambaut, A., and L. Bromham. 1998. Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution* **15**:442-448.

Rambaut, A., and M. Charleston. 2002. Phylogenetic Tree Editor and Manipulator v1.0 alpha 10. Department of Zoology, University of Oxford, Oxford, UK.

Reisz, R. R., and J. Müller. 2004. Molecular timescales and the fossil record: a paleontological perspective. *TRENDS in Genetics* **20**:237-241.

Rodríguez-Trelles, F., R. Tarrío, and F. J. Ayala. 2002. A methodological bias toward overestimation of molecular evolutionary time scales. *Proceedings of the National Academy of Sciences USA* **99**:8112-8115.

Rutschmann, F. 2005. Bayesian molecular dating using paml/multidivtime. A step-by-step manual. Version 1.5 (July 2005). Institute of Systematic Botany, University of Zurich, Zurich, Switzerland. Available from http://www.plant.ch.

Rutschmann, F., T. Eriksson, J. Schönenberger, and E. Conti. 2004. Did Crypteroniaceae really disperse out of India? Molecular dating evidence from *rbc*L, *ndh*F, and *rpl*16 sequences. *International Journal of Plant Sciences* **165**:69-83.

Sanderson, M. J. 1997. A nonparameteric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* **14**:1218-1231.

Sanderson, M. J. 1998. Estimating rate and time in molecular phylogenies: beyond the molecular clock? Pp. 242-264 *in* D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. Molecular systematics of plants II; DNA sequencing. Kluwer Academic Publishers, Norwell, MA.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a Penalized Likelihood approach. *Molecular Biology and Evolution* **19**:101-109.

Sanderson, M. J. 2003a. Molecular data from 27 proteins do not support a Precambrian origin of land plants. *American Journal of Botany* **90**:954-956.

Sanderson, M. J. 2003b. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**:301-302.

Sanderson, M. J., and J. A. Doyle. 2001. Sources of error and confidence intervals in estimating the age of Angiosperms from *rbc*L and 18S rDNA data. *American Journal of Botany* **88**:1499-1516.

Sanderson, M. J., J. L. Thorne, N. Wikström, and K. Bremer. 2004. Molecular evidence on plant divergence times. *American Journal of Botany* **91**:1656-1665.

Smith, A. B., and J. J. Peterson. 2002. Dating the time of origin of major clades: molecular clocks and the fossil record. *Annual Review of Earth and Planetary Sciences* **30**:65-88.

Soltis, P. S., D. E. Soltis, V. Savolainen, P. R. Crane, and T. G. Barraclough. 2002. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and

evidence for molecular living fossils. *Proceedings of the National Academy of Sciences USA* **99**:4430-4435.

Springer, M. S., W. J. Murphy, E. Eizirik, and S. J. O'Brien. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proceedings of the National Academy of Sciences USA* **100**:1056-1061.

Swofford, D. L. 2001. PAUP* 4.0b10: Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer, Sunderland, MA.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407-514 *in* D. M. Hillis, C. Moritz and B. K. Mable, eds. Molecular Systematics. Sinauer Associates, Sunderland, Massachusetts.

Sytsma, K. J., A. Litt, M. L. Zjhra, J. C. Pires, M. Nepokroeff, E. Conti, J. Walker, and P. G. Wilson. 2004. Clades, clocks, and continents: historical and biogeographical analysis of Myrtaceae, Vochysiaceae, and relatives in the southern hemisphere. *International Journal of Plant Sciences* **165**:85-105.

Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**:599-607.

Takahata, N. 1987. On the overdispersed molecular clock. *Genetics* **116**:169-179.

Takezaki, N., A. Rzhetsky, and M. Nei. 1995. Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution* **12**:823-833.

Tavaré, S., C. R. Marshall, O. Will, C. Soligo, and R. D. Martin. 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature* **416**:726-729.

Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**:689-702.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**:1647-1657.

Uyenoyama, M. 1995. A generalized least-squares estimate for the origin of sporophytic self-incompatibility. *Genetics* **139**:975-992.

Vos, R. A., and A. Ø. Mooers. 2004. Reconstructing divergence times for supertrees. Pp. 281-299 *in* O. R. P. Bininda-Emonds, ed. Phylogenetic supertrees: Combining information to reveal the tree of life. Kluwer Academic, Dordrecht, The Netherlands.

Welch, J. J., and L. Bromham. 2005. Molecular dating when rates vary. *TRENDS in Ecology and Evolution* **20**:320-327.

Whittle, C. A., and M. O. Johnston. 2003. Broad-scale analysis contradicts the theory that generation time affects molecular evolutionary rates in plants. *Journal of Molecular Evolution* **56**:223-233.

Wikström, N., V. Savolainen, and M. W. Chase. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society B: Biological Sciences* **268**:2211-2220.

Wikström, N., V. Savolainen, and M. W. Chase. 2003. Angiosperm divergence times: congruence and incongruence between fossils and sequence divergence estimates. Pp. 142-165 *in* P. C. J. Donoghue and M. P. Smith, eds. Telling the evolutionary time: molecular clocks and the fossil record. Taylor & Francis, London, UK.

Wray, G. A. 2001. Dating branches on the Tree of Life using DNA. *Genome Biology* **3**:1.1-1.7.

Wray, G. A., J. S. Levinton, and L. S. Shapiro. 1996. Molecular evidence for deep precambrian divergences among metazoan phyla. *Science* **274**:568-573.

Wu, C.-I., and W.-H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences USA* **82**:1741-1745.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences CABIOS* **13**:555-556.

Yang, Z. 2004. A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times. *Acta Zoologica Sinica* **50**:645-656.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo method. *Molecular Biology and Evolution* **14**:717-724.

Yang, Z., and A. D. Yoder. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic Biology* **52**:705-716.

Yoder, A. D., and Z. H. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution* **17**:1081-1090.

Yoon, H. S., J. D. Hackett, C. Ciniglia, G. Pinto, and D. Bhattacharya. 2004. A molecular timeline for the origin of photosynthetic Eukaryotes. *Molecular Biology and Evolution* **21**:809-818.

Zuckerkandl, E., and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. Pp. 97-166 *in* V. Bryson and H. Vogel, eds. Evolving Genes and Proteins. Academic Press, New York.

Tables

Table 1. Molecular dating methods that use a **molecular clock and one global rate of substitution**. The specifications about ease of use and popularity represent only the author's personal view and are therefore highly subjective. 1) Except model parameters, priors, or calibration constraints, 2) SE: standard error; CI: 95% confidence interval; CrI: 95% Bayesian credibility interval, 3) Unix software also runs under Mac OS X, as this operating system bases on Darwin, an open source UNIX environment. 4) PAUP* (Swofford *et al.*, 2001), DNAMLK (part of the PHYLIP package; Felsenstein, 1993), baseml (part of the paml package; Yang, 1997), MrBayes (Huelsenbeck and Ronquist, 2001), BEAST (Drummond and Rambaut, 2003) etc. 5) If the user provides a tree topology in addition to the sequences, the optimization runs much faster.

#	1	2	3
Method	**Linear regression**	**Mean path length**	**Langley-Fitch**
Author(s)	Nei (1987); Li and Graur (1991)	Bremer and Gustafsson (1997)	Langley and Fitch (1974)
Software where it's implemented	-	PATH (Britton 2002)	r8s (Sanderson 2003)
Current version	-	-	1.7
Runs on operating system(s)	-	Unix[3)]/Linux with Gpc 1999	Unix[3)]/Linux
Optimization strategy	-	-	Maximum likelihood
Input data[1)]	distance matrix	phylogram (with bl)	phylogram (with bl)
Allows multiple calibration points	yes	no	yes
Accounts for multiple datasets/partitions	no	no	no
Provides error estimates (SE/CI/CrI)[2)]	yes, CI	yes, mean path length CI	yes, internal and bootstrap CI's
Ease of use	easy	easy	medium
Popularity according to literature	very popular until about 10 years ago	popular	less popular

Table 1, continued...

#	**4**
Method	**ML with clock**
Author(s)	Felsenstein (1981)
Software where it's implemented	many phylogenetic packages[4)]
Current version	-
Runs on operating system(s)	depends on program
Optimization strategy	Maximum likelihood
Input data[1)]	sequence data (+ tree topology[5)])
Allows multiple calibration points	depends on program
Accounts for multiple datasets/partitions	no
Provides error estimates (SE/CI/CrI)[2)]	depends on program
Ease of use	depends on program
Popularity according to literature	very popular

Table 2. Molecular dating methods that that **correct for rate heterogeneity**. The specifications about ease of use and popularity represent only the author's personal view and are therefore highly subjective. 1) Except model parameters, priors, or calibration constraints. 2) SE: standard error; CI: 95% confidence interval; CrI: 95% Bayesian credibility interval. 3) Unix software also runs under Mac OS X, as this operating system bases on Darwin, an open source UNIX environment.

#	**5**
Method	**Linearized trees**
Author(s)	Li and Tanimura (1987)
Software where it's implemented	-
Current version	-
Runs on operating system(s)	-
Optimization strategy	depends on clock method
Input data[1)]	phylogram (with bl)
Allows multiple calibration points	depends on clock method
Accounts for multiple datasets/partitions	no
Provides error estimates (SE/CI/CrI)[2)]	depends on clock method
Ease of use	easy
Popularity according to literature	less popular

Table 2, continued...

#	**6**		
Method	**Local molecular clock**		
Author(s)	Hasegawa *et al.* (1989); Uyenoyama (1995); Rambaut and Bromham (1998); Yoder and Yang (2000)		
Software where it's implemented	baseml (paml; Yang, 1997)	QDate (Rambaut and Bromham, 1998)	Rhino (Rambaut, 2001)
Current version	3.14	1.1	1.2
Runs on operating system(s)	Unix[3)]/Linux/Windows	Mac OS 9.x/Unix[3)]/Linux/Windows	Mac OS 9.x/Unix[3)]/Linux/Windows
Optimization strategy	Maximum likelihood	Maximum likelihood	Maximum likelihood
Input data[1)]	phylogram (with bl)	sequence data + quartet definition	sequence data + tree topology
Allows multiple calibration points	yes	yes	yes
Accounts for multiple datasets/partitions	yes	no	only codon position partitions
Provides error estimates (SE/CI/CrI)[2)]	yes, SE	yes, CI	yes, CI
Ease of use	medium	easy	medium
Popularity according to literature	popular	less popular	popular, but more for rate comparisons

Table 2, continued...

#	**6**, continued...	
Method	**Local molecular clock**	
Author(s)	Hasegawa *et al.* (1989); Uyenoyama (1995); Rambaut and Bromham (1998); Yoder and Yang (2000)	
Software where it's implemented	r8s (Sanderson, 2003b)	BEAST (Drummond and Rambaut, 2003)
Current version	1.7	1.3
Runs on operating system(s)	Unix[3)]/Linux	Unix[3)]/Linux/Windows, requires Java 1.4
Optimization strategy	smoothing / minimizing optimality function	Bayesian MCMC
Input data[1)]	phylogram (with bl)	sequence data
Allows multiple calibration points	yes	yes
Accounts for multiple datasets/partitions	no	yes
Provides error estimates (SE/CI/CrI)[2)]	yes, CI, but separate bootstrapping is required	yes, CrI
Ease of use	medium	medium
Popularity according to literature	popular, but more for NPRS and PL methods	becoming increasingly popular

Table 3 (see next page). Molecular dating methods that that **incorporate rate heterogeneity**. The specifications about ease of use and popularity represent only the author's personal view and are therefore highly subjective

1) Except model parameters, priors, or calibration constraints.
2) SE: standard error; CI: 95% confidence interval; CrI: 95% Bayesian credibility interval.
3) Unix software also runs under Mac OS X, as this operating system bases on Darwin, an open source UNIX environment.
4) Plus one hyperparameter (autocorrelation value ν).
5) Plus several hyperparameters (describing speciation and extinction rate).
6) Implements a range of relaxed clock models by Drummond *et al.* (2006), but also the models by Thorne and Kishino (2002) and Aris-Brosou and Yang (2002).
7) Additionally, two graphical user interfaces called BEAUti and TRACER facilitate data setup and output analysis.
8) The method is implemented in a c program, but it is not meant for public access (as there is no documentation).
9) For defining the age of calibration points, different prior distributions are available, such as normal, lognormal, exponential, or gamma.

Table 3, continued**...**

#	**7**		**8**
Method	**NPRS**		**PL**
Author(s)	Sanderson (1997)		Sanderson (2002)
Software where it's implemented	r8s (Sanderson, 2003b)	TreeEdit (Rambaut and Charleston, 2002)	r8s (Sanderson, 2003b)
Current version	1.7	1.0a10	1.7
Runs on operating system(s)	Unix[3)]/Linux	Mac OS 8.6 or later, including Mac OS X	Unix[3)]/Linux
Optimization strategy	smoothing / minimizing optimality function	smoothing / minimizing optimality function	smoothing / minimizing optimality function
Input data1)	phylogram (with bl)	phylogram (with bl)	phylogram (with bl)
Model of rate evolution	rate autocorrelation	rate autocorrelation	rate autocorrelation
Allows multiple calibration points	yes	no	yes
Accounts for multiple datasets/partitions	no	no	no
Provides error estimates (SE/CI/CrI)[2)]	yes, internal and bootstrap CI's	no	yes, CI, but separate bootstrapping is required
Accounts for phylogenetic uncertainty	no	no	no
Ease of use	medium	easy (graphical user interface)	medium
Popularity according to literature	popular	popular	becoming increasingly popular

Table 3, continued**...**

#	9	10	11
Method	**AHRS**	**multidivtime**	**Phybayes**
Author(s)	Yang (2004)	Thorne *et al.* (1998), Kishino *et al.* (2001)	Aris-Brosou and Yang (2002)
Software where it's implemented	baseml (paml; Yang, 1997)	multidivtime (Thorne and Kishino, 2002)	Phybayes (Aris-Brosou and Yang, 2001)
Current version	3.14 (since 3.14beta5)	9/25/03	0.2e
Runs on operating system(s)	Unix[3)]/Linux/Windows	Unix[3)]/Linux/Windows	Unix[3)]/Linux/Windows
Optimization strategy	Maximum likelihood	Bayesian MCMC	Bayesian MCMC
Input data[1)]	sequence data + tree topology	sequence data + tree topology	sequence data + tree topology
Model of rate evolution	rate autocorrelation	rate autocorrelation	rate autocorrelation
Rates are drawn from	-	lognormal distribution	six different distributions
Prior distribution of divergence time	-	described as dirichlet distribution[4)]	described as generalized birth-death process[5)]
Allows multiple calibration points	yes	yes, as user-specified intervals	no
Accounts for multiple datasets/partitions	yes	yes	no
Provides error estimates (SE/CI/CrI)[2)]	yes , CI	yes, CrI	yes, CrI, but must be calculated by the user
Accounts for phylogenetic uncertainty	no	no, but accepts polytomies in input tree	no
Ease of use	medium	medium (use step-by-step manual; Rutschmann, 2005)	medium
Popularity according to literature	not yet very popular	becoming increasingly popular	not yet very popular

Table 3, continued...

#	12	13	14
Method	**variable rate models in BEAST**	**overdispersed clock**	**compound poisson**
Author(s)	Drummond *et al.* (2006)	Cutler (2000)	Huelsenbeck *et al.* (2000)
Software where it's implemented	BEAST (Drummond *et al.*, 2006)	c program (Cutler, 2000)	c program (Huelsenbeck *et al.*, 2000)
Current version	1.3	dating5.c	- [8]
Runs on operating system(s)	Unix[3]/Linux/Windows, requires Java 1.4	Unix[3]/Linux/Windows	- [8]
Optimization strategy	Bayesian MCMC	Maximum likelihood	Bayesian MCMC
Input data[1]	sequence data	phylogram (with bl)	- [8]
Model of rate evolution	various models implemented[6]	doubly stochastic poisson process	compound poisson process
Rates are drawn from	different distributions, such as log or lognormal	-	- [8]
Prior distribution of divergence time	no specific description, priors are uniform[9]	-	- [8]
Allows multiple calibration points	yes	yes	- [8]
Accounts for multiple datasets/partitions	yes	no	- [8]
Provides error estimates (SE/CI/CrI)[2]	yes, CrI	yes, CI	- [8]
Accounts for phylogenetic uncertainty	yes	no	- [8]
Ease of use	medium, a range of tutorials is available[7]	medium	- [8]
Popularity according to literature	becoming increasingly popular	not yet very popular	- [8]

Chapter II.

Did Crypteroniaceae really disperse out-of-India? Molecular dating evidence from *rbc*L, *ndh*F, and *rpl16* intron sequences

Frank Rutschmann, Torsten Eriksson[1], Jürg Schönenberger[2], and Elena Conti[3]

[1]*Bergius Foundation, Royal Swedish Academy of Sciences, SE-104 05 Stockholm, Sweden*

[2]*Department of Botany, Stockholm University, Lilla Frescativägen 5, SE-106 91 Stockholm, Sweden*

[3]*Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland*

Published in
International Journal of Plant Sciences 165 (4 Suppl.): S69-S83. 2004.

Abstract

Biogeographical and paleontological studies suggested that some ancient Gondwanan taxa have been carried by the rafting Indian plate from Gondwana to Asia. During this journey, the Indian island experienced dramatic latitudinal and climatic changes that caused massive extinctions in its biota. However, some taxa survived these conditions and dispersed "out-of-India" into South and South East Asia, after India collided with the Asian continent in the Early Tertiary. To test this hypothesis, independent estimates for lineage ages are needed. A published *rbc*L tree supported the sister group relationship between the South and South East Asian Crypteroniaceae (comprising *Crypteronia*, *Axinandra* and *Dactylocladus*) and a clade formed by the African Oliniaceae, Penaeaceae, and Rhynchocalycaceae and the Central and South American Alzateaceae. Molecular dating estimates suggested that Crypteroniaceae split from their West Gondwanan sister clade in the Early to Middle Cretaceous, and reached Asia rafting on the Indian plate. Here we present molecular evidence from additional chloroplast DNA regions and more taxa to test the validity of the out-of-India hypothesis for Crypteroniaceae. Both clock-based (Langley-Fitch) and clock-independent age estimates (NPRS and Penalized Likelihood), based on maximum likelihood analyses of three chloroplast DNA regions (*rbc*L, *ndh*F, and *rpl*16 intron), were used to infer the age of Crypteroniaceae. Our dating results suggest an ancient Gondwanan origin of Crypteroniaceae in the Early to Middle Cretaceous, followed by diversification on the Indian plate in the Early Tertiary and subsequent dispersal to South East Asia. These findings are congruent with recent molecular, paleontological, and biogeographic results in vertebrates. Within the biogeographic context of this study, we explore the critical assignment of paleobotanic and geological constraints to calibrate ultrametric trees.

Key words: Molecular dating, molecular clock, r8s, rates of substitution, penalized likelihood, nprs, clock calibration, biogeography, Gondwana, vicariance, Crypteroniaceae, Myrtales.

Introduction

Crypteroniaceae sensu stricto (Myrtales; Candolle 1857) are a small group of evergreen tropical shrubs and trees comprising three genera: *Crypteronia* Bl., with seven species, is the genus with the broadest distribution in South East Asia, including the Malay Peninsula, Sumatra, Java, Borneo, Philippines, Thailand, Vietnam, Myanmar, and New Guinea; *Dactylocladus* Oliv. has only one species, *Dactylocladus stenostachys*, endemic to Borneo; *Axinandra* Thw. includes one species, *Axinandra zeylanica*, endemic to Sri Lanka, and three other species with restricted distribution in the Malay peninsula and the northern part of Borneo (van Beusekom-Osinga and van Beusekom, 1975; Johnson and Briggs, 1984; Pereira and Wong, 1995; Conti *et al.*, 2002; Figure 1).

South East Asia and Sri Lanka are among the taxonomically most diverse regions on earth. In addition, they harbour high proportions of endemic species. For these reasons, both areas have been included among the 25 hotspots of biological diversity identified in a recent worldwide survey (Myers *et al.*, 2000). This remarkable species richness can be partially explained by the geological history of Sri Lanka and South East Asia. The uplift of the Himalayan chain caused by the collision of the Deccan plate (comprising India, Sri Lanka, and the Seychelles) with Laurasia during the Eocene (between 55 and 40 million years [mys] ago) and the generalized Late Tertiary aridification (Partridge, 1997; Willis and McElwain, 2002, pp. 197-198) led to an impoverishment of the tropical biome in Asia. Pockets of this biome, however, survived in refugial areas characterized by constant, tropical conditions, for example, in Sri Lanka and South East Asia. Only in these refugial areas, did tropical plants have a chance to survive the detrimental effects of Quaternary climate oscillations on the Indian subcontinent (Raven and Axelrod, 1974). The relictual nature of the South East Asian flora is also reflected in the great concentration of early diverging angiosperm clades in the fossil records of the subtropical forests of Asia-Australasia (Morley, 2001).

Crypteroniaceae had been proposed as being an ancient and relictual group on the basis of their distribution and morphology (van Vliet and Baas, 1975; van Beusekom-Osinga, 1977, p. 189). They represent an interesting case study to investigate the relative contributions of Laurasian and Gondwanan elements to the South Asian flora, because their members had been alternatively suggested as being of Laurasian or Gondwanan origin. For example, Meijer (1972) postulated a Gondwanan origin for *Axinandra*, a genus that he interpreted as being morphologically similar to the ancestor of the entire order Myrtales. Furthermore, Ashton and Gunatilleke (1987, p. 263), referring to the biogeographic history of *Axinandra*, stated: "The disjunct distribution and generalized morphology of this lowland rain forest genus suggest

considerable antiquity and possible spread into Asia by way of the Deccan Plate". The same authors suggested that *Axinandra* and other taxa were carried by the rafting Indian plate from Gondwana to Laurasia. After India collided with the Asian continent in the Early Tertiary, a few surviving Gondwanan elements dispersed "out-of-India" into South and South East Asia, which at the time lay in the same latitudinal and climatic zone (Morley, 2000). The out-of-India origin of Crypteroniaceae was also supported in recent biogeographic studies based on molecular dating estimates (Conti *et al.*, 2002; Morley and Dick, 2003).

The idea that splitting plates may carry biotic elements from one continent to the other had already been proposed by Axelrod (1971) and McKenna (1973). However, Raven and Axelrod (1974) noted that it is difficult to find evidence for out-of-India dispersal, because of the dramatic latitudinal and climatic changes that affected the Deccan plate during its journey from Gondwana to Laurasia and the ensuing massive extinctions in its biota. The same authors suggested a Laurasian origin for Crypteroniaceae (Raven and Axelrod, 1974). Recent molecular phylogenetic analyses of *rbc*L sequences in Myrtales (Conti *et al.*, 2002) supported that Crypteroniaceae form a monophyletic group comprising *Axinandra*, *Dactylocladus* and *Crypteronia* and identified a sister clade comprising: 1) Penaeaceae, a small group of 23 species in 7 genera endemic to the Cape Province of South Africa; 2) Oliniaceae, comprising a single genus with 8 species restricted to Eastern and Southern Africa; 3) Rhynchocalycaceae, with the single species *Rhynchocalyx lawsonioides*, a rare, evergreen tree endemic to Eastern Cape and KwaZulu-Natal in South Africa (Johnson and Briggs, 1984); and 4) Alzateaceae, with the single species *Alzatea verticillata*, a tree restricted to the submontane tropical forests of Bolivia, Peru, Panama, and Costa Rica (Graham, 1984) .

To investigate the biogeographic history of Crypteroniaceae, Conti *et al.* (2002) inferred the age of Crypteroniaceae by using three different molecular dating approaches applied to *rbc*L sequences. Because both phylogenetic relationships and dating estimates of relevant nodes were concordant with the geological history of the Deccan Plate in relation to West Gondwanan continents, the authors suggested a West Gondwanan origin for Crypteroniaceae and related families, with subsequent dispersal of Crypteroniaceae to the Asian continent via India. However, these conclusions were based on evidence from only one gene (*rbc*L) and limited taxon sampling from Crypteroniaceae and related families.

In this paper, we test the validity of previous conclusions on the out-of-India origin of Crypteroniaceae by expanding the taxon sampling to include four out of 12 described species of Crypteroniaceae and 13 out of 33 described species of their sister clade. Furthermore, we perform our analyses on DNA sequences of three chloroplast regions (*rbc*L, *ndh*F, *rpl*16 intron) and a combined dataset, compare the results of clock-dependent (Langley-Fitch, LF: Langley and Fitch,

1974) and clock-independent molecular dating methods (non-parametric rate smoothing, NPRS: Sanderson, 1997; and Penalized Likelihood, PL: Sanderson, 2002), and evaluate the level of error in our divergence time estimates by implementing a bootstrap approach (Baldwin and Sanderson, 1998; Sanderson and Doyle, 2001). We also discuss how problems of calibration in molecular dating analyses affect different conclusions on possible biogeographic scenarios for our study system.

Materials and Methods

Plant material and DNA extractions

For *Crypteronia paniculata*, *Crypteronia griffithii*, *Axinandra zeylanica*, *Dactylocladus stenostachys*, and *Olinia emarginata* we extracted total genomic DNA from silica dried leaf material. Leaf tissue was homogenized using glass beads and a MM 2000 shaker (Retsch GmbH, Haan, Germany). The DNA from these species was extracted with a method described in protocol D of Smith *et al.* (1991), which employs a 2% hexadecyl-trimethylammonium bromide (CTAB) extraction/lysis buffer. For all other taxa, the method of DNA extraction is given in Schönenberger and Conti (2003). Taxon names, voucher information, and GenBank accession numbers are listed in table 1.

PCR and DNA sequencing

Amplification and sequencing primers from Zurawski *et al.* (1981), Olmstead and Sweere (1994), and Baum *et al.* (1998) were used to generate DNA sequences of *rbc*L, *ndh*F and *rpl*16 intron, respectively. PCR amplifications were performed in a Biometra TGradient thermocycler, applying a thermal cycling program that consisted of 34 cycles of 0.5 min at 95 °C, 1 min at 49 °C to 52 °C, and 1.7 min at 72 °C, followed by a terminal extension of 10 min at 72 °C. In order to successfully detect amplified DNA target regions and possible contamination, PCR products were separated on 1% agarose gels, stained with ethidium bromide, and visualized under UV light. Successfully amplified PCR products were purified with the QIAquick PCR Purification Kit (QIAGEN, Basel, Switzerland). Cycle-sequencing reactions were performed using the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems, Applera Europe B.V., Rotkreuz, Switzerland). For a few taxa, we were unable to amplify the entire *rpl*16 intron; in these cases two additional internal primers, MF and MR, were used (Schönenberger and Conti, 2003). Cycle-sequencing reactions were performed in a GeneAmp PCR System 9700 (Applied Biosystems) by using a temperature cycle of 10 s at 96 °C, 5 s at 50 °C, and 4 min at 60 °C (25 cycles). The sequencing fragments were cleaned with MicroSpin G-50 columns (Amersham Pharmacia Biotech Europe GMBH, Dübendorf, Switzerland) to remove excess dye terminators before loading them on an ABI Prism 3100 Genetic Analyzer (Applied Biosystems).

The software Sequencher 3.1.1 (Gene Codes, Ann Arbor, MI, USA) was used to edit and assemble complementary strands. Base positions were individually double-checked for agreement between the complementary strands. *Rbc*L sequences were readily aligned by eye, while *ndh*F and *rpl*16 intron sequences were first aligned using Clustal X 1.81 (Thompson *et al.*, 1997) prior

to adjusting the alignments by eye in the software MacClade 4.0 (Maddison and Maddison, 2000). For the *rpl*16 intron dataset, the variable region between nucleotides 810 and 1031 was deleted, because we were unable to produce a reasonable alignment within that region. The datasets used for further phylogenetic analyses contained 24 taxa and 1280 (*rbc*L), 981 (*ndh*F), 1010 (*rpl*16 intron), and 3271 (all three datasets combined) aligned positions.

Phylogenetic analyses

Maximum Likelihood (ML) optimization was used to find the best tree for each of the three separate data partitions and for the combined data matrix, including all characters. A Neighbor-joining tree calculated under the JC69 substitution model (Jukes and Cantor, 1969) was used as the starting tree to estimate the optimal ML parameters under 56 different models of evolution in Modeltest 3.06 (Posada and Crandall, 1998). The best substitution model was selected by performing hierarchical likelihood ratio tests (Felsenstein, 1981; Huelsenbeck and Rannala, 1997). The selected optimal models were all submodels of the general time reversible (GTR) model (Rodríguez *et al.*, 1990). For the *rbc*L dataset, the K81uf+G+I model was selected (Kimura, 1981): unequal base frequencies (A=0.2673, C=0.1941, G=0.2488, T=0.2898), one transition rate (AG/CT: 1.4245), two transversion rates (AC/GT: 1, AT/CG: 0.3281), gamma distribution of rates among sites with alpha shape parameter 0.7280 (Yang, 1993), proportion of invariable sites 0.7098. For the *ndh*F dataset, the TVM+G model was selected: unequal base frequencies (A=0.3085, C=0.1481, G=0.1529, T=0.3905), one transition rate (AG/CT: 1.5243), four transversion rates (AC:1.2165, AT:0.1544, CG:1.4160, GT:1), gamma distribution of rates among sites with alpha shape parameter 0.3887, no proportion of invariable sites. For the *rpl*16 intron dataset, the K81uf+G model (Kimura, 1981) was selected: unequal base frequencies (A=0.3716, C=0.1651, G=0.1685, T=0.2948), one transition rate (AG/CT: 1.3446), two transversion rates (AC/GT: 1, AT/CG: 0.4837), gamma distribution of rates among sites with alpha shape parameter 0.8358, no proportion of invariable sites. For the combined dataset, the TVM+G+I model was selected: unequal base frequencies (A=0.314, C=0.1692, G=0.1912, T=0.3256), one transition rate (AG/CT: 1.5), four transversion rates (AC:1.1651, GT: 1, AT: 0.3469, CG: 0.7904), gamma distribution of rates among sites with alpha shape parameter 0.8288, proportion of invariable sites 0.4007.

The estimated parameters were then used in a ML heuristic search using 100 random addition sequences, tree bisection reconnection (TBR) branch swapping, and steepest descent activated, implemented in PAUP 4.0b10 (Swofford, 2001). The trees were rooted on the branch leading to *Mouriri helleri* (Memecylaceae) and four representatives of Melastomataceae, which were constrained to be monophyletic. These choices were justified by the results of more

inclusive phylogenetic analyses of Myrtales (Conti *et al.*, 1996 and 2002). Statistical support for each clade was estimated by generating 1000 bootstrap pseudoreplicates using the ML fast-heuristic search option in PAUP. All phylogenetic analyses were performed on a 2 Ghz Pentium IV machine under Red Hat Linux 8.0.

Molecular Dating

When performing molecular dating analyses, several crucial choices need to be made that might affect the estimated ages, including selection of molecular dating method, gene sampling, and calibration method. The first choice is to decide whether the analyses should be based on the assumption of rate constancy (molecular clock) or whether they should allow rates to vary across branches of a tree. To evaluate whether the sequences of each data partition evolved in a clock-like fashion, a likelihood ratio (LR) test was performed by comparing the scores of ML trees with and without a molecular clock enforced (Felsenstein, 1981; Sanderson, 1998, Nei and Kumar, 2000). To gain some insight into the relative performance of different molecular dating methods, we compared the results of the clock-dependent Langley-Fitch (LF; Langley and Fitch, 1974) and the clock-independent non-parametric rate smoothing (NPRS; Sanderson, 1997) and Penalized Likelihood (PL; Sanderson, 2002) analyses, as implemented in r8s 1.6 (Sanderson, 2003). Both latter methods relax the assumption of rate constancy by smoothing changes of substitution rates across the tree. NPRS is an entirely non-parametric method that estimates rates and times via a least-squares smoothing criterion, whereas PL is a semi-parametric technique that attempts to combine the statistical power of parametric methods with the robustness of non-parametric methods. Briefly, PL relies on a data-driven cross-validation procedure that sequentially prunes taxa from the tree, estimates parameters from the submatrix for a given smoothing value, predicts the data that were removed by using the estimated parameters, and calculates the χ^2 error associated with the difference between predicted and observed data of the removed submatrix. The optimal smoothing level corresponds to the lowest χ^2 error (Sanderson, 2002).

The optimal ML trees estimated in PAUP 4.0b10 for each data partition and for the combined dataset were saved with branch lengths (Figures 2-5a) and then used as input trees in r8s. To establish the position of the root in the basal branch of the ML trees, *Myrtus communis* and *Eugenia uniflora* (Myrtaceae) were used as dating outgroups (see Conti *et al.*, 1996; 1997). To evaluate the overall branch length from the root to a tip of a tree, it is necessary to know the lengths of the basal branches. In an additive tree, only the sum of these lengths is known, and the place where the root attaches to the basal branches is undefined (Figure 6a). Dating outgroup choice influences the position of the root attachment point, hence the lengths of the basal branches (Figure 6b) and the relative contribution of individual branches to the total paths from

the root to the tips (Sanderson and Doyle, 2001; Sanderson, 2002). Therefore, root position affects the calculation of absolute substitution rates and the smoothing of differential substitution rates across the tree.

After root position was established, the dating outgroup was removed in r8s prior to molecular dating. For the PL analyses, the optimal smoothing parameter, ranging between 0.001 to 1000, was selected prior to the dating by performing a cross-validation procedure. To calculate the absolute substitution rates across the tree, optimization via Truncated-Newton (TN) algorithm was chosen for the Langley-Fitch and PL methods, and the POWELL algorithm for the NPRS dating.

All age estimations in r8s were started five times to provide different starting conditions (a random set of initial divergence times), a practice aimed at preventing the optimization algorithms from converging to a local plateau. Age estimations were performed only for nodes A, B, and C, because these nodes are crucial to testing the out-of-India origin of Crypteroniaceae. Node A represents the diversification of the Crypteroniaceae crown group; node B the origin of the Crypteroniaceae stem lineage (equivalent to the time at which Crypteroniaceae split off from their West Gondwanan sister group); and node C represents the split between the South American *Alzatea* and its African sister clade (see Figures 2-5).

To evaluate statistical support for the estimated ages, we performed a bootstrap resampling procedure (Efron and Tibshirani, 1993). For all molecular dating analyses performed on the combined dataset, 100 bootstrap pseudoreplicates were generated using the program SEQBOOT from the Phylip package, version 3.6a3 (Felsenstein, 2002). While the topology of the optimal ML trees was kept fixed, branch lengths for each pseudoreplicate were estimated by ML with the selected substitution model in PAUP (Sanderson, 1997). With this approach, 100 bootstrap trees with the same topology, but different branch lengths were generated and individually analyzed in 100 molecular dating procedures as described above. For the PL analysis, the optimal smoothing parameter for each bootstrap replicate was calculated prior to the dating procedures. Using the r8s-bootstrap-kit (Eriksson, 2002), relative branch lengths from the 100 bootstrapped trees were transformed into a distribution of 100 absolute ages for each of nodes A, B, and C, respectively. After checking for normality, the obtained age distributions were used to calculate the mean, standard deviation, and 95% confidence interval of each age estimate (tables 2-4).

Calibration

To transform the resulting relative branch lengths into absolute ages for nodes A, B, and C (Figure 5b) it is necessary to fix or constrain a node to an absolute age. The calibration

procedure represents one of the most sensitive choices in molecular dating analyses (Yang and Yoder, 2003; Thorne and Kishino, 2002; Wikström *et al.*, 2001; Sanderson, 1998). Calibration can be performed by reference either to the fossil record (paleobotanic dating) or to known vicariance events (geological dating; Sanderson, 1998; Hillis *et al.*, 1996). Either approach can establish only the minimum age at the calibration point, most likely resulting in an underestimation of divergence times (Tavaré *et al.*, 2002). In the following section we consider the problems associated with each of the three calibration points that we selected for our analyses.

From an analytical point of view, the ideal calibration point would be as close as possible to the node to be estimated, in order to reduce potential sources of error in age estimation (Wikström *et al.*, 2001). In our tree, this was possible only with geological calibration (see below), because the fossil record of Crypteroniaceae is too uncertain. Heterocolpate pollen tentatively assigned to Crypteroniaceae from the Middle Miocene (Muller, 1975) is difficult to distinguish from heterocolpate pollen of Melastomataceae, Memecylaceae, Oliniaceae, Penaeaceae, and Rhynchocalycaceae (Morley and Dick, 2003). Therefore, we were forced to look for paleobotanic calibration points outside of Crypteroniaceae.

The phylogenetically closest fossils were in Melastomataceae. Renner *et al.* (2001) and Renner and Meyer (2001) used fossil seeds from the Miocene of central Europe (Collinson and Pingen, 1992) to constrain the origin of Melastomeae. However, the assignment of these seeds to Melastomeae is not straightforward. Collinson and Pingen (1992, p. 134) stated: "[These fossil seeds] are most similar to seeds of members of the tribes Osbeckieae [Melastomeae] and Rhexieae, but differ in several significant features, especially the presence of multicellular tubercles". Therefore, it is difficult to know whether these fossils should be assigned to the base of the Melastomeae crown group or to more recent nodes in the tribe. With these caveats in mind, we assigned a probably very conservative age of 26 mys (as suggested by Renner *et al.*, 2001) to node D, representing the crown group of Melastomeae in our current taxon sampling (Figures 2-5).

Fossil leaves from the Early Eocene of North Dakota (53 mys; Hickey, 1977) can also be used to constrain a node in Melastomataceae. However, the assignment of these macrofossils to a specific node is problematic. In his description of these fossil leaves, Hickey (1977) stated that they resemble most closely the leaves of extant Miconieae and Merianieae, but cautioned: "They all differ, however, in not being deeply cordate and in having tertiaries which do not form a good V pattern" (Hickey, 1977, p. 144). Renner *et al.* (2001) conservatively assigned these leaves, characterized by the acrodromous leaf venation typical of extant Melastomataceae, to the node that subtends the crown group of the entire Melastomataceae, including the basal Kibessieae. However, these fossil leaves can also be assigned to the crown group that includes Miconieae and

Micronieae, as the comments by Hickey (1977) might imply (see also Renner *et al.*, 2001). Our current sampling of Melastomataceae does not include representatives of the basal Kibessieae. However, also in light of the biases in the macrofossil record discussed by Morley and Dick (2003) it does not seem unreasonable to assign an age of 53 mys to the node that comprises our current sampling of Melastomataceae (node E, see Figures 2-5; see also Renner, 2004).

Several recent studies have used geological calibration points for molecular dating estimates, for example, in *Phylica* (Richardson *et al.*, 2001), Laurales (Renner *et al.*, 2000), ranid frogs (Bossuyt and Milinkovitch, 2001), and ratite birds (Cooper *et al.*, 2001). In our analyses, all ML trees from either separate or combined datasets supported the sister group relationship between the South American Alzateaceae and the African clade (see Results). This pattern, supported by a bootstrap value of 86% in the combined ML tree (see Figure 5a), represents a rather clear geological signature, and can be used as a calibration point, despite caveats of potential circularity. Therefore, we assigned an age of 90 mys, representing the final split between South America and Africa, to node C (see Figure 5b).

Results

Phylogenetic analyses

One single optimal ML tree was found for each dataset with a log-likelihood score of –lnL = 3408.68 (*rbc*L), 4119.76 (*ndh*F), 4603.61 (*rpl*16 intron), and 12460.2 (combined dataset; Figures 2-5a).

All optimal trees from the three individual and the combined datasets showed common results: 1) *Crypteronia*, *Axinandra*, and *Dactylocladus* (all Crypteroniaceae) formed a monophyletic group, with bootstrap support values (BS) between 98% and 100% (Figures 2-5a); 2) *Alzatea* (Alzateaceae), *Rhynchocalyx* (Rhynchocalycaceae), and all Penaeaceae and Oliniaceae included in this analysis formed another clade with BS between 51% and 86%; 3) These two clades were sister to each other with BS between 67% and 99%; and 4) *Alzatea* was sister to the clade formed by *Rhynchocalyx*, Oliniaceae, and Penaeaceae with BS between 79 and 97%. In all these clades, the highest bootstrap support values were obtained in the combined analysis.

Molecular dating

By enforcing the molecular clock in Paup 4.0b10 we obtained an optimal ML tree for each dataset with log-likelihood scores of –lnL = 3429.36 (*rbc*L), 4169.04 (*ndh*F), 4648.36 (*rpl*16 intron), and 12531.27 (combined dataset). Comparisons between clock and non-clock trees by applying likelihood ratio (LR) tests rejected clock-like evolution for all datasets (LR = 41.35, *rbc*L; 98.56, *ndh*F; 89.52, *rpl*16 intron; 142.12, combined dataset; degrees of freedom = 22, confidence interval = 95%). The results of molecular dating analyses using both clock-dependent and clock-independent approaches for the three separate and for the combined datasets are summarized in tables 2-4, and the PL chronogram for the combined dataset is shown in Figure 5b. Smoothing parameter values of 0.1 (*rbc*L), 0.01 (*ndh*F), 0.001 (*rpl*16 intron), and 0.00316 (combined dataset), selected via a cross-validation procedure in r8s, were used for Penalized Likelihood age estimations.

Discussion

The topologies of all optimal ML trees from the three individual and the combined datasets were congruent with phylogenies published by Clausing and Renner (2001), Conti *et al.* (2002), Schönenberger and Conti (2003), and Sytsma *et al.* (this issue). However, the trees differed slightly in the detailed topological resolution within the clades mentioned in the Results section and in the branch lengths.

Comparisons among dating methods

Comparisons among the nodal ages estimated by the three dating methods showed remarkable differences (table 2-4), depending on the methods themselves, but also on the position of the calibration node within the tree. A discrepancy between clock-based and clock-independent age estimates was expected, because likelihood ratio tests strongly rejected the assumption of rate constancy for all datasets. Differences in the rates of nucleotide substitution between branches in a tree are also known as lineage effects (Britten, 1986; Gillespie, 1991).

In general, the non-parametric rate smoothing method (NPRS; Sanderson, 1997), which relaxes the assumption of rate constancy by smoothing changes of substitution rates across the tree, consistently produced the highest rate differences between the branches (as visualized in the ratograms produced by r8s; data not shown), thus the ages estimated using NPRS were either much younger or older than those obtained by using Langley-Fitch (LF: Langley and Fitch, 1974) – depending on the position of the calibration node and data partition. This is because NPRS tends to overfit the data, thus causing rapid rate fluctuations in certain regions of a tree (Sanderson, 2002). The semi-parametric Penalized Likelihood method (PL; Sanderson, 2002) tries to alleviate this problem by selecting the optimal smoothing parameter via a data-driven cross-validation procedure (Green and Silverman, 1994). The application of PL resulted in rates of nucleotide substitution as well as age estimates which were for most branches between those calculated with LF and NPRS.

By calibrating the trees at nodes E or D (tables 2 and 3), the ages estimated for nodes A, B, and C using NPRS were consistently older than those produced using PL, whereas the ages obtained using LF were younger than the PL results. The likely explanation for this effect lies in the two long branches below node E (Figures 2-5a). Depending on the dating method, different rates of nucleotide substitution are assigned to these branches (ratograms not shown), producing considerably different absolute ages at nodes located on the other side of the root of the tree.

Comparisons among DNA regions

By calibrating the trees at nodes E or D (tables 2 and 3), the ages for nodes A, B, and C estimated using the *ndh*F dataset were generally much older than those based on the other two datasets. Reciprocally, when we calibrated the trees at node C (table 4), the ages for nodes E and D were much younger in the *ndh*F based analysis than by using the *rbc*L or *rpl*16-intron datasets. Nodal ages estimated from *rbc*L and *rpl*16 intron sequences were similar to each other. Comparisons of the three ML phylograms show that the branches below node E are significantly longer in the *ndh*F phylogram (Figure 3) than the same branches in the *rbc*L and *rpl*16-intron phylograms (Figures 2 and 4).

The phenomenon of striking differences in the tempo and mode of evolution between different genes is well known, but its effects on divergence time estimation are poorly understood (Goremykin *et al.*, 1996; Sanderson and Doyle, 2001). In the present paper, we can only speculate on possible explanations for the anomalous results obtained from *ndh*F sequences and suggest research directions that might prove fruitful to investigate the role of gene-specific effects on molecular dating estimates.

For example, the bias of nucleotide substitutions in both coding and non-coding sequences of the plant chloroplast genome is strongly dependent on the composition of the two flanking bases (Morton, 1997a, 1997b). One possible explanation for the older ages obtained from *ndh*F sequences might lie in the differential influence that the two neighboring bases could have on the substitution type of a certain nucleotide in our *ndh*F sequences as compared to *rbc*L and *rpl*16 intron sequences. It is also reasonable to ask whether the occurrence of an *ndh*F pseudogene might explain gene-specific effects on age estimates, as *ndh*F pseudogenes have been reported, for example, in orchids (Neyland and Urbatsch, 1996). However, translation of *ndh*F sequences into the corresponding amino acids did not reveal the presence of any stop codons, and multiple alignment of *ndh*F sequences required only gaps in multiples of three, suggesting that our *ndh*F sequences likely represent functional gene copies. Furthermore, PCR amplifications using an *ndh*F-specific primer pair (Olmstead and Sweere, 1994) did not reveal any PCR products of different lengths, and direct sequencing of double-stranded PCR products produced unequivocal electropherograms, characterized by single peaks at all positions. Finally, the topology of the optimal ML tree based *ndh*F was congruent to the other trees based on the *rbc*L and *rpl*16 intron sequences. Nevertheless, we cannot exclude the possibility that an *ndh*F pseudogene might exist in some or all of the studied taxa, perhaps influencing the molecular evolutionary behavior of the functional *ndh*F gene copies that we likely sequenced for this study (Bromham and Penny, 2003). Another potential explanation for the different dating results obtained from *ndh*F sequences might be sought in alignment effects. However, experiments using

a modified *ndh*F dataset from which all gapped regions were removed prior to analysis produced age estimates similar to those obtained with the gapped dataset (data not shown). Additional theoretical and experimental studies of gene-specific effects on molecular dating estimates are clearly needed (Bromham and Penny, 2003).

Although nested likelihood ratio tests of the three separate datasets indicated that the three chloroplast regions used in our dating analyses evolved according to different models and parameters of nucleotide substitutions, we proceeded to estimate nodal ages from the combined data matrix, because we wished to compare results from the latter with those from the three separate sequence matrices. The exceedingly older or younger nodal ages estimated from *ndh*F sequences, depending on the position of the calibration point, seemed to further justify dataset combination, for it has been suggested that the combination of sequences with different evolutionary patterns might compensate for unusual patterns in any single DNA region (Wikström *et al.*, 2001; Qiu *et al.*, 1999).

Congruence between geological and biological history

One of the major goals of biogeographic studies is to elucidate the historical genesis of current plant distributions. In an evolutionary framework, it is assumed that geological events of the past, for example the emergence and/or the elimination of major geographic barriers to range expansion, likely left a mark on the phylogenetic and biogeographic history of biotic elements (Lieberman, 2000). Therefore, to support the hypothesis that geological events shaped the current distribution of any taxa, one would need to demonstrate congruence between geological and biological history both in terms of pattern and time (table 5). At the level of pattern, one would expect correspondence between the sequence of geological events and the sequence of cladogenetic events. Paleogeological reconstructions and the topology of phylogenetic trees provide the necessary evidence for pattern congruence. At the level of time, the specific timing of geological events must be compatible with the timing of cladogenetic events (nodal ages) inferred from molecular or other dating methods. Both lines of evidence (pattern and time) are necessary, but independently not sufficient, to support a key role of geology in shaping current taxic distributions. If evidence for congruence between geology and biology can be produced at the levels of both pattern and time, there is no need to invoke other types of explanatory processes for current biotic distributions (table 5; see also Sober, 1988; Hunn and Upchurch, 2001).

Geological events that influence biological distributions include plate fragmentation, as in the classic interpretation of vicariance, or the expansion of a lineage due to the temporary elimination or reduction of a geographic barrier, followed by the emergence of a new barrier producing vicariant sister groups, as in the recently proposed concept of geodispersal (Lieberman,

1997; Lieberman, 2000). In the next section we will discuss whether the phylogenetic relationships and molecular dating estimates of Crypteroniaceae warrant a key role for geological events in explaining the current distribution of this group and its sister clade.

Congruence between geology and biology for the out-of-India hypothesis of Crypteroniaceae

After comparing the results of different dating methods and datasets (see above), it seemed most reasonable to use the ML tree topology (Figure 5a) and the PL ages (Figure 5b) calculated from the combined data matrix to reconstruct the biogeographic history of Crypteroniaceae. A previous phylogenetic and molecular dating study based exclusively on *rbc*L sequences proposed an ancient Gondwanan origin for Crypteroniaceae in the Early to Middle Cretaceous, followed by dispersal to the Deccan plate (comprising Madagascar, India, Sri Lanka, and the Seychelles) as it was rafting along the African coast, and subsequent dispersal from India to South East Asia after collision of the Indian plate with Asia in the Middle Eocene (Conti *et al.*, 2002). Is this biogeographic reconstruction congruent with both pattern and timing of cladogenetic events (table 5), as estimated from the phylogenetic and molecular dating analyses of the expanded datasets used in the present study?

The combined ML tree (Figure 5a) strongly supports (BS = 99%) the sister group relationship between the South East Asian Crypteroniaceae and the West Gondwanan clade and the split between the South American Alzateaceae and the African clade (BS= 86%). Therefore, at the level of pattern, the sequence of cladogenetic events is congruent with the sequence of geological events, if we consider that the Deccan Plate rafted along the African coast between the Lower and Middle Cretaceous (Scotese *et al.*, 1988; Morley, 2000), with likely island chain connections between the two plates up to the Early Maastrichtian (Morley, 2000), and that separation between Africa and South America was completed by approximately 90 mys (McLoughlin, 2001), although trans-oceanic dispersal routes between Africa and South America likely existed between 84 and 65 mys (McDougal and Douglas, 1988; Hallam, 1994; Morley, 2000).

At the level of time, results are more controversial. The deviations in age estimates due to the use of different calibrations (tables 2-4) indicate that calibration is one of the most critical issues in molecular phylogenetic dating. The ages for the origin of Crypteroniaceae (node B), obtained from PL optimization on the combined ML tree, ranged from a minimum value of 62 mys, estimated by fixing node D to 26 mys, to a higher value of 101 mys, estimated by fixing node E to 53 mys, and a maximum value of 109 mys, estimated by fixing node C to 90 mys (see tables 2-4 and Figure 5b). As explained in Methods, the assignments of fossil seeds from the

Miocene of central Europe (Collinson and Pingen, 1992) to node D (Melastomeae crown group) and fossil leaves from the Eocene of North Dakota (Hickey, 1977) to node E (Melastomataceae crown group) most likely represent large underestimations of nodal ages. Furthermore, Morley and Dick (2003) extensively reviewed the fossil record for Melastomataceae and argued that its abrupt appearance at northern temperate latitudes during the Eocene and Miocene may simply reflect colonization from ancient Gondwanan lineages. Given these considerations, it seems more plausible to suggest an origin of the Crypteroniaceae stem lineage that is closer to the older ages(101-106 mys) obtained with our three calibration points. Which biogeographic scenario is congruent with this interpretation for the age of Crypteroniaceae?

According to paleogeographic reconstructions, East Gondwana - including India - split from West Gondwana between 165 and 150 mys ago (Krutzsch, 1989; McLoughlin, 2001; Briggs, 2003). Therefore, a traditional vicariant explanation for the origin of Crypteroniaceae - with overland dispersal from West to East Gondwana, followed by tectonic split - is incompatible with our dating estimates for the age of angiosperms (190-140 mys; Sanderson and Doyle, 2001; Wikström *et al.*, 2001). It is more probable that the biogeographic history of Crypteroniaceae might reflect a temporary reduction or even elimination of the oceanic barrier between Africa and the Deccan Plate (at that time comprising Madagascar, India, Sri Lanka, and the Seychelles Plateau), as the plate drifted northward along the African coast for a rather extended period of time (over 40 mys) between the Early and Late Cretaceous (Morley, 2000; Scotese *et al.*, 1988; Briggs, 2003). It has also been suggested that small islands or land bridges between West Gondwana and the Deccan Plate facilitated short- to medium-distance dispersal over the Mozambique Channel of other biotic elements, including some groups of dinosaurs, crocodiles, mammals (Krause *et al.*, 1999; Krause and Maas, 1990), frogs (Biju and Bossuyt, 2003; Bossuyt and Milinkovitch, 2001), lizards, snakes, turtles, and caecilians (Briggs, 2003). Ashton and Gunatilleke (1987) suggested that the total distance between West Gondwana and the Deccan plate (still connected to Madagascar) remained more or less constant (about 420 km) until approximately 84 mys ago, when the plate separated from Madagascar and started to drift northwards (Storey *et al.*, 1995; Plummer and Belle, 1995; McLoughlin, 2001). Pollen records suggested that plant dispersal from Africa to Madagascar and the Indian plate continued on a regular basis, presumably until the middle Maastrichtian (65-71 mys ago; Morley and Dick, 2003). Therefore, India's role in the biogeographic history of Crypteroniaceae most likely did not conform to a purely vicariant pattern, involving direct dispersal prior to barrier formation (Wiley, 1988; Morrone and Crisci, 1995), but rather to the dynamics of range expansion following barrier reduction (geodispersal; Lieberman, 2000; see also Stace, 1989).

Extinction played a prominent role in the history of the ancient Gondwanan elements of India's biotas, as India traveled rapidly across latitudes during the Middle to Late Cretaceous (McLoughlin, 2001; Morley, 2000). Its biotas were affected by massive volcanism at the Cretaceous-Tertiary boundary (approximately 65 mys ago; Officer *et al.*, 1987), extensive aridification during the Late Tertiary (following the uplift of the Himalayan chain caused by India's collision with Southern Asia between 55 and 49 mys ago; Beck *et al.*, 1995), and further cycles of aridity associated with glaciations during the Quaternary (Raven and Axelrod, 1974; Bande and Prakash, 1986; Ashton and Gunatilleke, 1987; Morley, 2000). *Axinandra zeylanica* is endemic in Sri Lanka which was probably connected to India until 6000 years ago (McLoughlin, 2001). Southwestern India together with Sri Lanka served as refugial areas, where some ancient Gondwanan taxa escaped extinction (Raven and Axelrod, 1974; Guleria, 1992; Morley, 2000). Some of these relictual taxa dispersed to South East Asia, where *Crypteronia* sp., *Dactylocladus stenostachys* and the other three species of *Axinandra* occur to this day. South East Asia has also long been recognized as a refugium where the equable oceanic conditions allowed tropical lineages to survive (Bande and Prakash, 1986; Morley, 2000; Takhtajan, 1987).

To summarize, our current phylogenetic and molecular dating results from expanded taxic and genetic sampling suggest a possible congruence between biological and geological history that is compatible with a central role played by the Deccan Plate in transporting the stem lineage of Crypteroniaceae from West Gondwana to Asia, most likely in a time frame comprised between the Middle to Late Cretaceous. However, our results remain open to debate, especially in light of the difficult assignment of paleobotanic and geological constraints to specific nodes in the phylogeny. It is our hope that the addition of more fossil calibration points, further taxonomic sampling from Crypteroniaceae and additional groups, and the use of dating methods that allow for multiple, contemporary constraints on the phylogeny will allow us to refine our interpretations of the biogeographic history of Crypteroniaceae and related clades.

References

Arbogast, B. S., S. V. Edwards, J. Wakeley, P. Beerli, and J. B. Slwinski. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology, Evolution and Systematics* **33**:707-740.

Aris-Brosou, S. 2001. Phybayes: a program for phylogenetic analyses in a Bayesian framework. Department of Biology (Galton Laboratory), University College London, London, UK.

Aris-Brosou, S., and Z. Yang. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA Phylogeny. *Systematic Biology* **51**:703-714.

Aris-Brosou, S., and Z. Yang. 2003. Bayesian models of episodic evolution support a Late Precambrian explosive diversification of Metazoa. *Molecular Biology and Evolution* **20**:1947-1954.

Baldwin, B. G., and M. J. Sanderson. 1998. Age and rate of diversification of the hawaiian silversword alliance (Compositae). *Proceedings of the National Academy of Sciences USA* **95**:9402-9406.

Bell, C. D., D. E. Soltis, and P. S. Soltis. 2005. The age of the Angiosperms: A molecular timescale without a clock. *Evolution* **59**:1245-1258.

Berry, P. E., W. J. Hahn, K. J. Sytsma, J. C. Hall, and A. Mast. 2004. Phylogenetic relationships and biogeography of *Fuchsia* (Onagraceae) based on noncoding nuclear and chloroplast DNA data. *American Journal of Botany* **94**:601-614.

Bremer, K., E. M. Friis, and B. Bremer. 2004. Molecular phylogenetic dating of asterid flowering plants shows early Cretaceous diversification. *Systematic Biology* **53**:496-505.

Bremer, K., and M. H. G. Gustafsson. 1997. East Gondwana ancestry of the sunflower alliance of families. *Proceedings of the National Academy of Sciences USA* **94**:9188-9190.

Britten, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**:1393-1398.

Britton, T. 2005. Estimating divergence times in phylogenetic trees without a molecular clock. *Systematic Biology* **54**:500-507.

Britton, T., B. Oxelman, A. Vinnersten, and K. Bremer. 2002. Phylogenetic dating with conficence intervals using mean path lengths. *Molecular Phylogenetics and Evolution* **24**:58-65.

Bromham, L., and D. Penny. 2003. The modern molecular clock. *Nature Reviews Genetics* **4**:216-224.

Bromham, L., A. Rambaut, R. Fortey, A. Cooper, and D. Penny. 1998. Testing the Cambrian explosion hypothesis by using a molecular dating technique. *Proceedings of the National Academy of Sciences USA* **95**:12386-12389.

Bromham, L., and M. Woolfit. 2004. Explosive radiations and the reliability of molecular clocks: island endemic radiations as a test case. *Systematic Biology* **53**:758-766.

Bromham, L. D., A. Rambaut, M. D. Hendy, and D. Penny. 2000. The power of relative rates test depends on the data. *Journal of Molecular Evolution* **50**:296-301.

Bryant, D., C. Semple, and M. Steel. 2004. Supertree methods for ancestral divergence dates and other applications. Pp. 129-150 *in* O. R. P. Bininda-Emonds, ed. Phylogenetic supertrees: Combining information to reveal the tree of life. Kluwer Academic, Dordrecht, The Netherlands.

Chaw, S. M., C. C. Chang, H. L. Chen, and W.-H. Li. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *Journal of Molecular Evolution* **58**:424-441.

Conti, E., T. Eriksson, J. Schönenberger, K. J. Sytsma, and D. A. Baum. 2002. Early Tertiary out-of-India dispersal of Crypteroniaceae: evidence from phylogeny and molecular dating. *Evolution* **56**:1931-1942.

Conti, E., F. Rutschmann, T. Eriksson, K. J. Sytsma, and D. A. Baum. 2004. Calibration of molecular clocks and the biogeographic history of Crypteroniaceae: A reply to Moyle. *Evolution* **58**:1874-1876.

Cutler, D. J. 2000. Estimating divergence times in the presence of an overdispersed molecular clock. *Molecular Biology and Evolution* **17**:1647-1660.

Douzery, E. J. P., E. A. Snell, E. Bapteste, F. Delsuc, and H. Philippe. 2004. The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proceedings of the National Academy of Sciences USA* **101**:15386-15391.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**:e88.

Drummond, A. J., and A. Rambaut. 2003. BEAST v1.0. Available at http://evolve.zoo.ox.ac.uk/beast/.

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, New York.

Eriksson, T. 2002. The r8s-bootstrap-kit. Bergianska University, Stockholm, Sweden.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368-376.

Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22**:521-265.

Felsenstein, J. 1993. Phylogenetic Inference Package (PHYLIP), version 3.5. University of Washington, Seattle, WA.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Inc., Sunderland, MA.

Foote, M., J. P. Hunter, C. M. Janis, and J. J. Sepkoski Jr. 1999. Evolutionary and preservational constraints on origins of biologic groups: divergence times of Eutherian mammals. *Science* **283**:1310-1314.

Foote, M., and J. J. Sepkoski Jr. 1999. Absolute measures of the completeness of the fossil record. *Nature* **398**:415-417.

Gillespie, J. H. 1986. Rates of molecular evolution. *Annual Review of Ecology, Evolution and Systematics* **17**:637-665.

Gillespie, J. H. 1991. The causes of molecular evolution. Oxford University Press, Oxford, GB.

Gillooly, J. F., A. P. Allen, G. B. West, and J. H. Brown. 2005. The rate of DNA evolution: Effects of body size and temperature on the molecular clock. *Proceedings of the National Academy of Sciences USA* **102**:140-145.

Graur, D., and W. Martin. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *TRENDS in Genetics* **20**:80-86.

Hasegawa, M., H. Kishino, and T. Yano. 1989. Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. *Journal of Human Evolution* **18**:461-476.

Hasegawa, M., J. L. Thorne, and H. Kishino. 2003. Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes & Genetic Systems* **78**:267-283.

Heckman, D. S., D. M. Geiser, B. R. Eidell, R. L. Stauffer, N. L. Kardos, and S. B. Hedges. 2001. Molecular Evidence for the Early Colonization of Land by Fungi and Plants. *Science* **293**:1129-1133.

Hedges, S. B., and S. Kumar. 2003. Genomic clocks and evolutionary timescales. *TRENDS in Genetics* **19**:200-206.

Hedges, S. B., and S. Kumar. 2004. Precision of molecular time estimates. *TRENDS in Genetics* **20**:242-247.

Hedges, S. B., P. H. Parker, C. G. Sibley, and S. Kumar. 1996. Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**:226-229.

Hillis, D. M., B. K. Mable, and C. Moritz. 1996. Applications of molecular systematics: the state of the field an a look to the future. Pp. 515-543 *in* D. M. Hillis, C. Moritz and B. K. Mable, eds. Molecular Systematics. Sinauer, Sunderland, MA.

Ho, S. Y. W., M. J. Phillips, A. Cooper, and A. J. Drummond. 2005a. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular Biology and Evolution* **22**:1561-1568.

Ho, S. Y. W., M. J. Phillips, A. J. Drummond, and A. Cooper. 2005b. Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. *Molecular Biology and Evolution* **22**:1355-1363.

Huelsenbeck, J. P., B. Larget, and D. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* **154**:1879-1892.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* **17**:754.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution* **170-179**:170-179.

Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* **18**:352-361.

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**:1789-1796.

Langley, C. H., and W. Fitch. 1974. An estimation of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution* **3**:161-177.

Li, W.-H. 1997. Molecular Evolution. Sinauer Press, Sunderland MA, USA.

Li, W.-H., and D. Graur. 1991. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA.

Li, W.-H., and M. Tanimura. 1987. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326**:93-96.

Linder, H. P., C. R. Hardy, and F. Rutschmann. 2005. Taxon sampling effects in molecular clock dating: an example from the African Restionaceae. *Molecular Phylogenetics and Evolution* **35**:569-582.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. Debry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. De Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**:610-614.

Magallón, S. A. 2004. Dating lineages: molecular and paleontological approaches to the temporal framework of clades. *International Journal of Plant Sciences* **165**:7-21.

Magallón, S. A., and M. J. Sanderson. 2001. Absolute diversification rates in Angiosperm clades. *Evolution* **55**:1762-1780.

Martin, A. P., and S. R. Palumbi. 1993. Body size, metabolic rate, generation time and the molecular clock. *Proceedings of the National Academy of Sciences USA* **90**:4087-4091.

Muse, S. V., and B. S. Weir. 1992. Testing for equality of evoltuionary rates. *Genetics* **132**:269-276.

Near, T. J., and M. J. Sanderson. 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **359**:1477-1483.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Ohta, T. 2002. Near-neutrality in evolution of genes and in gene regulation. *Proceedings of the National Academy of Sciences USA* **99**:16134-16137.

Ota, R., and D. Penny. 2003. Estimating changes in mutational mechanisms of evolution. *Journal of Molecular Evolution* **57**:S233–S240.

Pérez-Losada, M., J. T. Høeg, and K. A. Crandall. 2004. Unraveling the evolutionary radiation of the Thoracian Barnacles using molecular and morphological evidence: a comparison of several divergence time estimation approaches. *Systematic Biology* **53**:244-264.

Rambaut, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**:395-399.

Rambaut, A. 2001. Rhino v1.1. Available at http://evolve.zoo.ox.ac.uk/.

Rambaut, A., and L. Bromham. 1998. Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution* **15**:442-448.

Rambaut, A., and M. Charleston. 2002. Phylogenetic Tree Editor and Manipulator v1.0 alpha 10. Department of Zoology, University of Oxford, Oxford, UK.

Reisz, R. R., and J. Müller. 2004. Molecular timescales and the fossil record: a paleontological perspective. *TRENDS in Genetics* **20**:237-241.

Rodríguez-Trelles, F., R. Tarrío, and F. J. Ayala. 2002. A methodological bias toward overestimation of molecular evolutionary time scales. *Proceedings of the National Academy of Sciences USA* **99**:8112-8115.

Rutschmann, F. 2005. Bayesian molecular dating using paml/multidivtime. A step-by-step manual. Version 1.5 (July 2005). Institute of Systematic Botany, University of Zurich, Zurich, Switzerland. Available from http://www.plant.ch.

Rutschmann, F., T. Eriksson, J. Schönenberger, and E. Conti. 2004. Did Crypteroniaceae really disperse out of India? Molecular dating evidence from *rbc*L, *ndh*F, and *rpl*16 sequences. *International Journal of Plant Sciences* **165**:69-83.

Sanderson, M. J. 1997. A nonparameteric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* **14**:1218-1231.

Sanderson, M. J. 1998. Estimating rate and time in molecular phylogenies: beyond the molecular clock? Pp. 242-264 *in* D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. Molecular systematics of plants II; DNA sequencing. Kluwer Academic Publishers, Norwell, MA.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a Penalized Likelihood approach. *Molecular Biology and Evolution* **19**:101-109.

Sanderson, M. J. 2003a. Molecular data from 27 proteins do not support a Precambrian origin of land plants. *American Journal of Botany* **90**:954-956.

Sanderson, M. J. 2003b. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**:301-302.

Sanderson, M. J., and J. A. Doyle. 2001. Sources of error and confidence intervals in estimating the age of Angiosperms from *rbc*L and 18S rDNA data. *American Journal of Botany* **88**:1499-1516.

Sanderson, M. J., J. L. Thorne, N. Wikström, and K. Bremer. 2004. Molecular evidence on plant divergence times. *American Journal of Botany* **91**:1656-1665.

Smith, A. B., and J. J. Peterson. 2002. Dating the time of origin of major clades: molecular clocks and the fossil record. *Annual Review of Earth and Planetary Sciences* **30**:65-88.

Soltis, P. S., D. E. Soltis, V. Savolainen, P. R. Crane, and T. G. Barraclough. 2002. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and

evidence for molecular living fossils. *Proceedings of the National Academy of Sciences USA* **99**:4430-4435.

Springer, M. S., W. J. Murphy, E. Eizirik, and S. J. O'Brien. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proceedings of the National Academy of Sciences USA* **100**:1056-1061.

Swofford, D. L. 2001. PAUP* 4.0b10: Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer, Sunderland, MA.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407-514 *in* D. M. Hillis, C. Moritz and B. K. Mable, eds. Molecular Systematics. Sinauer Associates, Sunderland, Massachusetts.

Sytsma, K. J., A. Litt, M. L. Zjhra, J. C. Pires, M. Nepokroeff, E. Conti, J. Walker, and P. G. Wilson. 2004. Clades, clocks, and continents: historical and biogeographical analysis of Myrtaceae, Vochysiaceae, and relatives in the southern hemisphere. *International Journal of Plant Sciences* **165**:85-105.

Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**:599-607.

Takahata, N. 1987. On the overdispersed molecular clock. *Genetics* **116**:169-179.

Takezaki, N., A. Rzhetsky, and M. Nei. 1995. Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution* **12**:823-833.

Tavaré, S., C. R. Marshall, O. Will, C. Soligo, and R. D. Martin. 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature* **416**:726-729.

Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**:689-702.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**:1647-1657.

Uyenoyama, M. 1995. A generalized least-squares estimate for the origin of sporophytic self-incompatibility. *Genetics* **139**:975-992.

Vos, R. A., and A. Ø. Mooers. 2004. Reconstructing divergence times for supertrees. Pp. 281-299 *in* O. R. P. Bininda-Emonds, ed. Phylogenetic supertrees: Combining information to reveal the tree of life. Kluwer Academic, Dordrecht, The Netherlands.

Welch, J. J., and L. Bromham. 2005. Molecular dating when rates vary. *TRENDS in Ecology and Evolution* **20**:320-327.

Whittle, C. A., and M. O. Johnston. 2003. Broad-scale analysis contradicts the theory that generation time affects molecular evolutionary rates in plants. *Journal of Molecular Evolution* **56**:223-233.

Wikström, N., V. Savolainen, and M. W. Chase. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society B: Biological Sciences* **268**:2211-2220.

Wikström, N., V. Savolainen, and M. W. Chase. 2003. Angiosperm divergence times: congruence and incongruence between fossils and sequence divergence estimates. Pp. 142-165 *in* P. C. J. Donoghue and M. P. Smith, eds. Telling the evolutionary time: molecular clocks and the fossil record. Taylor & Francis, London, UK.

Wray, G. A. 2001. Dating branches on the Tree of Life using DNA. *Genome Biology* **3**:1.1-1.7.

Wray, G. A., J. S. Levinton, and L. S. Shapiro. 1996. Molecular evidence for deep precambrian divergences among metazoan phyla. *Science* **274**:568-573.

Wu, C.-I., and W.-H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences USA* **82**:1741-1745.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences CABIOS* **13**:555-556.

Yang, Z. 2004. A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times. *Acta Zoologica Sinica* **50**:645-656.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo method. *Molecular Biology and Evolution* **14**:717-724.

Yang, Z., and A. D. Yoder. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic Biology* **52**:705-716.

Yoder, A. D., and Z. H. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution* **17**:1081-1090.

Yoon, H. S., J. D. Hackett, C. Ciniglia, G. Pinto, and D. Bhattacharya. 2004. A molecular timeline for the origin of photosynthetic Eukaryotes. *Molecular Biology and Evolution* **21**:809-818.

Zuckerkandl, E., and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. Pp. 97-166 *in* V. Bryson and H. Vogel, eds. Evolving Genes and Proteins. Academic Press, New York.

Tables

Table 1. Species names, sources, and GenBank accession numbers of the DNA sequences used in the analyses. Herbaria acronyms: Z=Zurich, BOL=Bolus (University of Cape Town). Superscript numbers refer to sources in published articles. References: 1) Schönenberger and Conti, 2003; 2) Conti *et al.*, 1996; 3) Conti *et al.*, 2002; 4) Clausing and Renner, 2001; 5) Renner *et al.*, 2001; 6) Renner and Meyer, 2001.

Taxon	**Voucher or source**	**GenBank accession numbers**		
		***rbc*L**	***ndh*F**	***rpl*16 intron**
Alzatea verticillata Ruiz & Pavon	[1) 2) 4)]	U26316[2)]	AF215591[4)]	AY151598[1)]
Axinandra zeylanica Thwaites	Peter Ashton, s.n., Sri Lanka	AY078157[3)]	AJ605094	AJ605107
Brachysiphon acutus (Thunb.) A. Juss.	J. Schönenberger 365 (Z), (BOL)	AJ605084	AJ605095	AY151605[1)]
Brachysiphon fucatus (L.) Gilg	J. Schönenberger 357 (Z), (BOL	AJ605085	AJ605096	AY151606[1)]
Brachysiphon microphyllus Rourke	J. Schönenberger 386 (Z), (BOL)	AJ605086	AJ605097	AY151608[1)]
Crypteronia griffithii C.B. Clarke	Shawn Lum s.n., Singapore	AJ605087	AJ605098	AJ605108
Crypteronia paniculata Blume	Peter Ashton s.n., Brunei	AY078153[3)]	AJ605099	AY151597[1)]
Dactylocladus stenostachys Oliver	Peter Becker, s.n., Brunei	AY078156[3)]	AJ605100	AJ605109
Endonema retzioides A. Juss	J. Schönenberger 370 (Z), (BOL)	AJ605088	AJ605101	AY151611[1)]
Eugenia uniflora L.	[4)]	AF294255[4)]	AF215592[4)]	AF215627[4)]
Medinilla humbertiana Gaudich.	[4)]	AF215517[4)]	AF215557[4)]	AF215602[4)]
Mouriri helleri Aublet	[4) 5)]	AF270752[4)]	AF322230[5)]	AF215611[4)]
Myrtus communis L.	[4)]	AF294254[4)]	AF215593[4)]	AF215628[4)]
Olinia emarginata Davy	J. Schönenberger 579, cultivated, Kirstenbosch Botanical Garden, (Z)	AJ605089	AJ605102	AY151601[1)]

Table 1, continued...

Olinia ventosa (L.) Cuf.	[1) 4)]	AF215546[4)]	AF215594[4)]	AY151604[1)]
Osbeckia chinensis L.	[4)]	AF215525[4)]	AF215570[4)]	AF210378[4)]
Penaea mucronata L.	[1) 3) 4)]	AJ605090	AF270756[4)]	AY151620[1)]
Rhexia virginica L.	[2) 4)]	U26334[2)]	AF215587[4)]	AF215623[4)]
Rhynchocalyx lawsonioides Oliver	[1) 2) 4)]	U26336[2)]	AF270757[4)]	AY151599[1)]
Saltera sarcorolla (L.) Bullock	J. Schönenberger 360 (Z), (BOL)	AJ605091	AJ605103	AY151621[1)]
Sonderothamnus petraeus (Barker f.) R. Dahlgren	J. Schönenberger 362 (Z), (BOL)	AY078154[3)]	AJ605104	AY151622[1)]
Stylapterus ericoides A. Juss. ssp. *pallidus* R. Dahlgren	J. Schönenberger 355 (Z), (BOL)	AJ605092	AJ605105	AY151625[1)]
Stylapterus micranthus R. Dahlgren	M. Johns s.n. (Z)	AJ605093	AJ605106	AY151627[1)]
Tibouchina urvilleana (DC.) Cogn.	[2) 6)]	U26339[2)]	AF272820[6)]	AF322234[6)]

Table 2. Ages in million years (mys) estimated for nodes A, B, and C based on *rbc*L, *ndh*F, and *rpl*16 intron sequences and three different methods implemented in the r8s software (Sanderson, 2002). All trees were calibrated at node E with an age of 53 mys. μ: mean, σ: standard deviation, CI: 95% confidence interval, nnd.: bootstrapped ages not normally distributed, n.a.: data not available because node D is not present in the *rbc*L dataset due to different tree topology.

	***rbc*L**	***ndh*F**	***rpl*16 intron**	**combined dataset**
Langley-Fitch (LF)				
A	31.51	41.64	23.85	29.44, μ=29.91, σ=3.21, CI: 29.28-30.55
B	80.22	71.93	49.03	58.89, μ=58.16, σ=5.16, CI: 57.13-59.18
C	64.46	49.21	37.1	53.00, μ=43.63, σ=4.35, CI: 42.76-44.49
D	n.a.	45.43	42.4	42.69, μ=42.31, σ=2.08, CI: 41.9-42.72
NPRS				
A	57.08	153.7	56.21	82.68, μ=85.24, σ=11.93, CI: 82.87-87.61
B	111.44	217.3	106	152.64, μ=149.12, σ=16.51, CI: 145.84-152.39
C	100.56	196.1	94.76	135.68, μ=133.75, σ=15.84, CI: 130.6-136.89
D	n.a.	42.4	40.15	38.16, μ=38.35, nnd.
PL				
optimal smoothing parameter	α=0.1	α=0.01	α=0.001	α=0.00316
A	39.37	92.75	43.06	53.00, μ=54.26, σ=7.58, CI: 52.76-55.76
B	96.91	154.58	91.09	100.7, μ=103.28, σ=12.32, CI: 100.84-105.73
C	80.26	128.08	76.19	83.03, μ=85.91, σ=12.45, CI: 83.44-88.38
D	n.a.	44.17	39.75	42.4, μ=40.93, σ=2.18, CI: 40.5-41.36

Table 3. Ages in million years (mys) estimated for nodes A, B, and C based on *ndh*F and *rpl*16 intron sequences and three different methods implemented in the r8s software (Sanderson, 2002). All trees were calibrated at node D with an age of 26 mys. μ: mean, σ: standard deviation, CI: 95% confidence interval, nnd.: bootstrapped ages not normally distributed. Ages for the *rbc*L dataset are not available because node D is not present in this dataset due to different tree topology.

	***ndh*F**	***rpl*16 intron**	**combined dataset**
Langley-Fitch (LF)			
A	23.85	14.63	17.93, μ=18.44, nnd.
B	41.17	30.06	35.86, μ=35.82, σ=3.59, CI: 35.11-36.53
C	28.17	22.75	26, μ=26.87, σ=2.92, CI: 26.29-27.44
D	30.33	32.5	32.28, μ=32.65, σ=1.64, CI: 32.32-32.98
NPRS			
A	94.25	36.4	56.33, μ=58.11, σ=9.39, CI: 56.25-59.98
B	133.25	68.64	104, μ=101.53, σ=12.96, CI: 98.56-104.1
C	120.25	61.36	92.44, μ=91.07, σ=12.27, CI: 88.63-93.5
D	32.5	34.32	36.11, μ=31.78, nnd.
PL			
optimal smoothing parameter	α=0.01	α=0.001	α=0.00316
A	54.6	28.17	32.5, μ=34.6, σ=5.37, CI: 33.54-35.67
B	91	59.58	61.75, μ=65.79, σ=8.54, CI: 64.1-67.49
C	75.4	49.83	50.92, μ=54.73, σ=8.35, CI: 53.07-56.38
D	31.2	34.67	32.5, μ=33.77, σ=1.82, CI: 33.41-34.13

Table 4. Ages in million years (mys) estimated for nodes A, B, and C based on *rbc*L, *ndh*F, and *rpl*16 intron sequences and three different methods implemented in the r8s software (Sanderson, 2002). All trees were calibrated at node C with an age of 90 mys. μ: mean, σ: standard deviation, CI: 95% confidence interval, nnd.: bootstrapped ages not normally distributed, n.a.: data not available because node D is not present in the *rbc*L dataset due to different tree topology.

	***rbc*L**	***ndh*F**	***rpl*16 intron**	**combined dataset**
Langley-Fitch (LF)				
A	44	76.15	57.86	62.07, μ=62.08, σ=7.35, CI: 60.63-63.54
B	112	131.54	118.93	124.14, μ=120.52, σ=10.25, CI: 118.49-125.60
C	n.a.	83.08	102.86	90, μ=88.12, nnd.
D	74	96.92	128.57	111.72, μ=110.42, σ=11.09, CI: 108.22-112.63
NPRS				
A	51.08	70.54	53.39	54.84, μ=57.6, σ=6.97, CI: 56.21-58.98
B	99.73	99.73	100.68	101.25, μ=100.5, σ=4.23, CI: 99.66-101.34
C	n.a.	19.46	38.14	25.31, μ=26.15, σ=3.45, CI: 25.46-26.83
D	47.43	24.32	50.34	35.16, μ=36.15, σ=4.19, CI: 35.32-36.98
PL				
optimal smoothing parameter	α=0.1	α=0.01	α=0.001	α=0.00316
A	44.15	65.17	50.87	57.45, μ=57.31, σ=7.2, CI: 55.88-58.74
B	108.68	108.62	107.61	109.15, μ=108.89, σ=8.52, CI: 107.2-110.58
C	n.a.	31.03	46.96	46.96, μ=43.79, σ=7.02, CI: 42.4-45.18
D	59.43	37.24	62.61	57.45, μ=56.71, σ=8.39, CI: 55.04-58.37

Table 5. Relationships between geology and biology at the levels of pattern and time. Correspondence at both levels is necessary to support a geodispersalist origin (sensu Lieberman, 2000) of current biotic distributions. See Discussion for further explanations.

	Geology	**Biology**
Pattern	Sequence of geologic events	Sequence of cladogenetic events
Time	Timing of geologic events	Timing of cladogenetic events

Figures

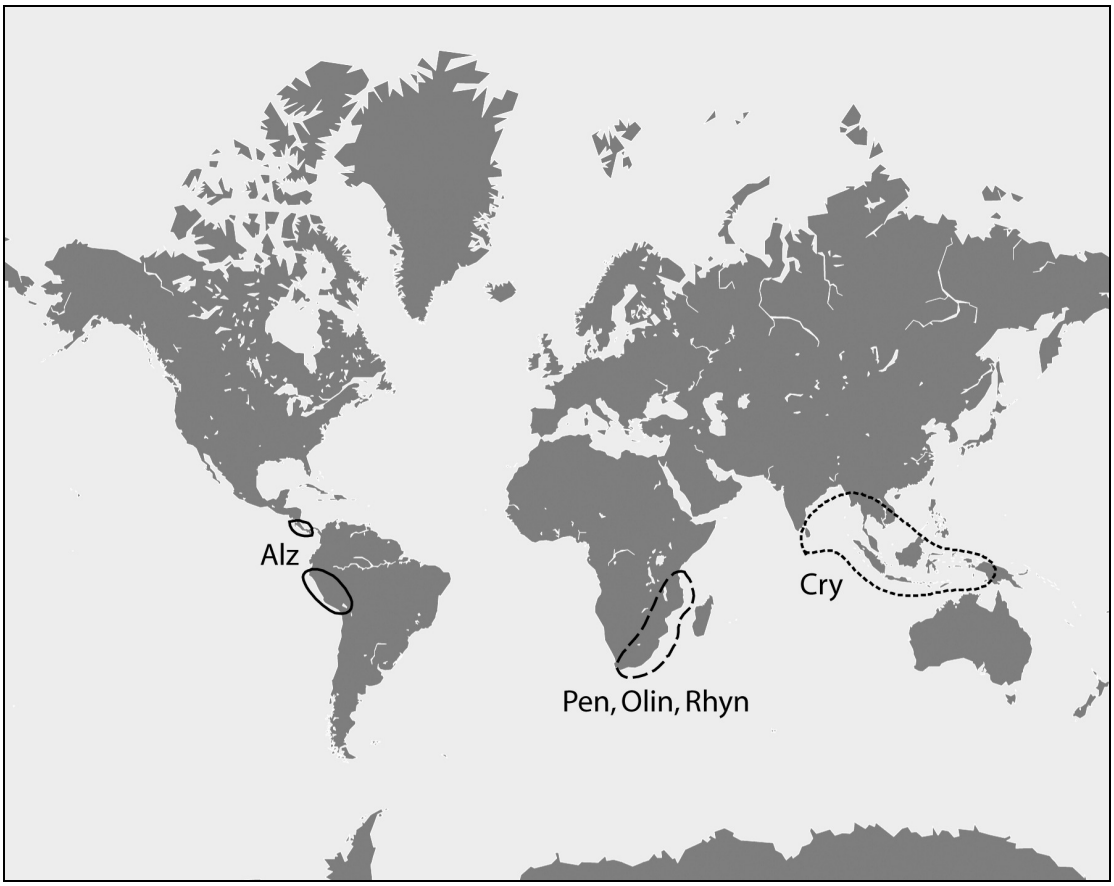


Figure 1. Current distribution of Crypteroniaceae and related taxa. **Cry**: Crypteroniaceae, **Alz**: Alzateaceae, **Pen**: Penaeaceae, **Olin**: Oliniaceae, **Rhyn**: Rhynchocalycaceae.

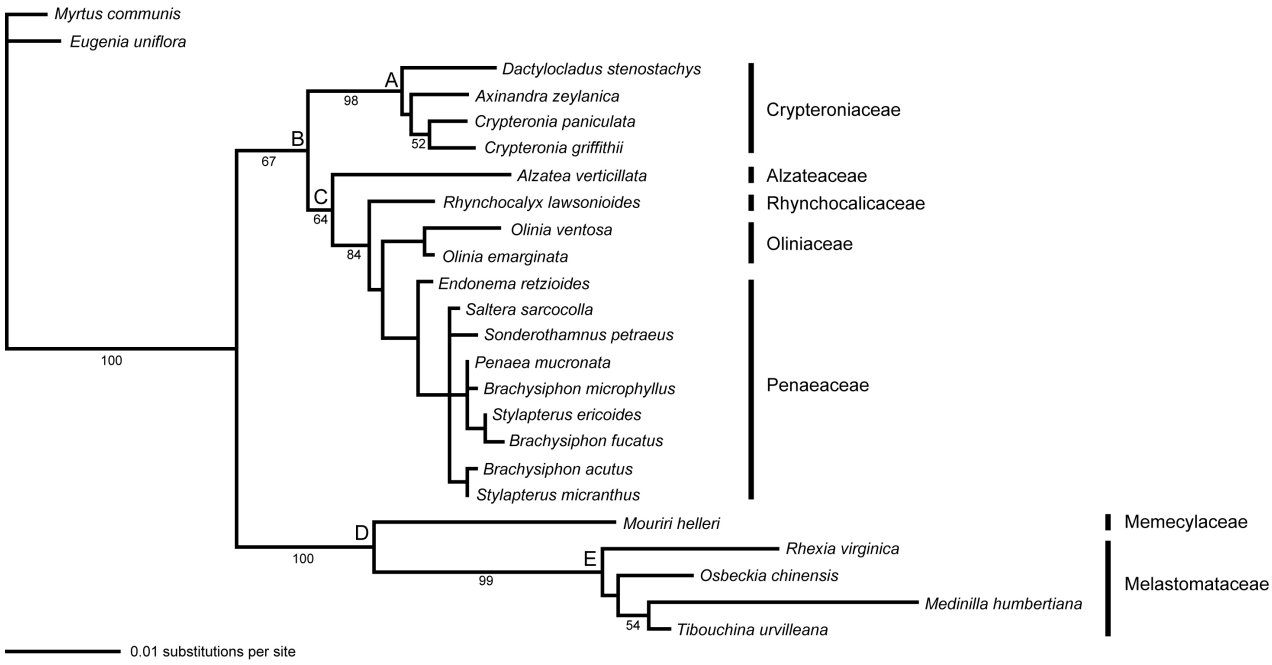


Figure 2. Maximum Likelihood (ML) tree, based on the *rbc*L dataset (1280 characters). Bootstrap support values are reported below the branches. Nodes of interest: A (diversification of Crypteroniaceae crown group), B (origin of Crypteroniaceae stem lineage), C (diversification of the West Gondwanan crown group), D (crown group of Melastomeae), and E (diversification of Melastomataceae crown group).

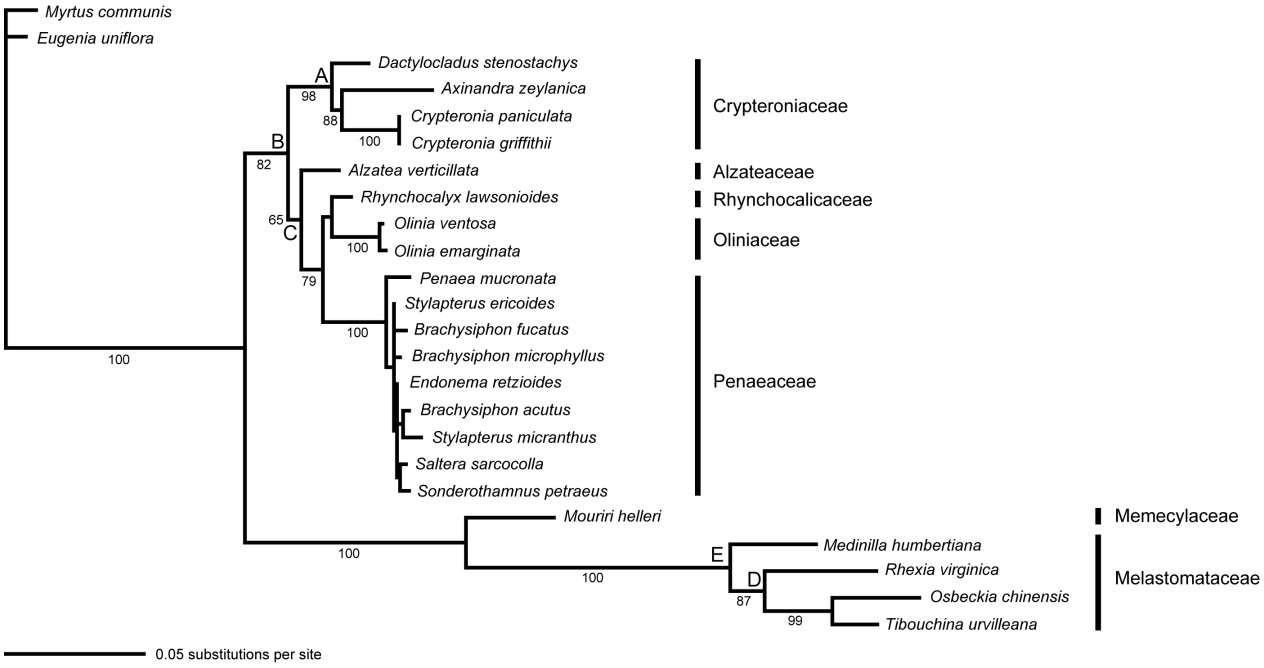


Figure 3. Maximum Likelihood (ML) tree, based on the *ndh*F dataset (981 characters). Bootstrap support values and description of nodes as in legend of Figure 2.

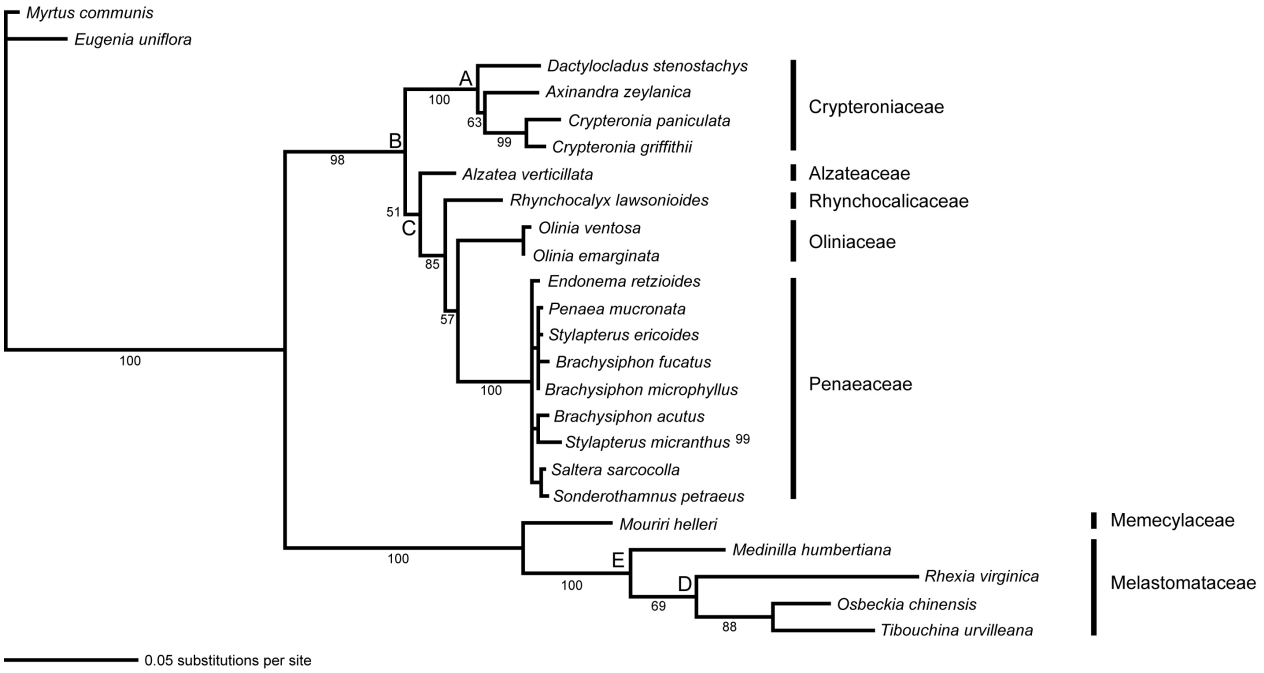


Figure 4. Maximum Likelihood (ML) tree, based on the *rpl*16 intron dataset (1010 characters). Bootstrap support values and description of nodes as in legend of Figure 2.

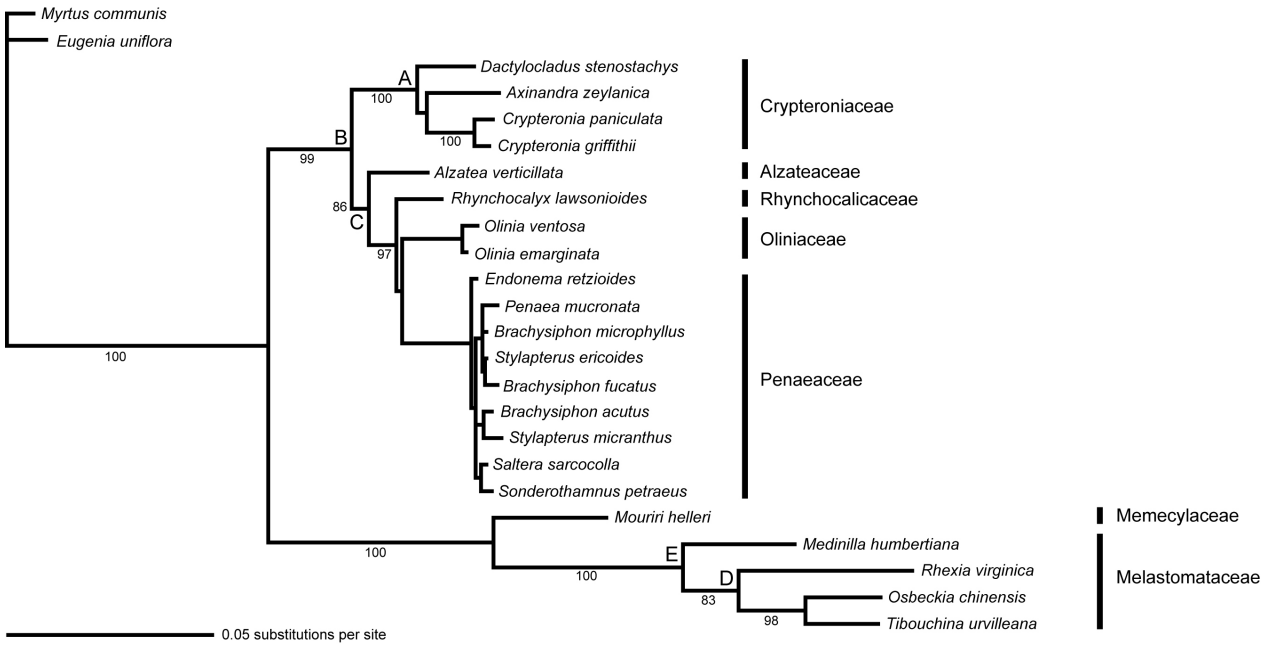


Figure 5a. Maximum Likelihood (ML) tree, based on the combined dataset (3271 characters). Bootstrap support values and description of nodes as in legend of Figure 2.

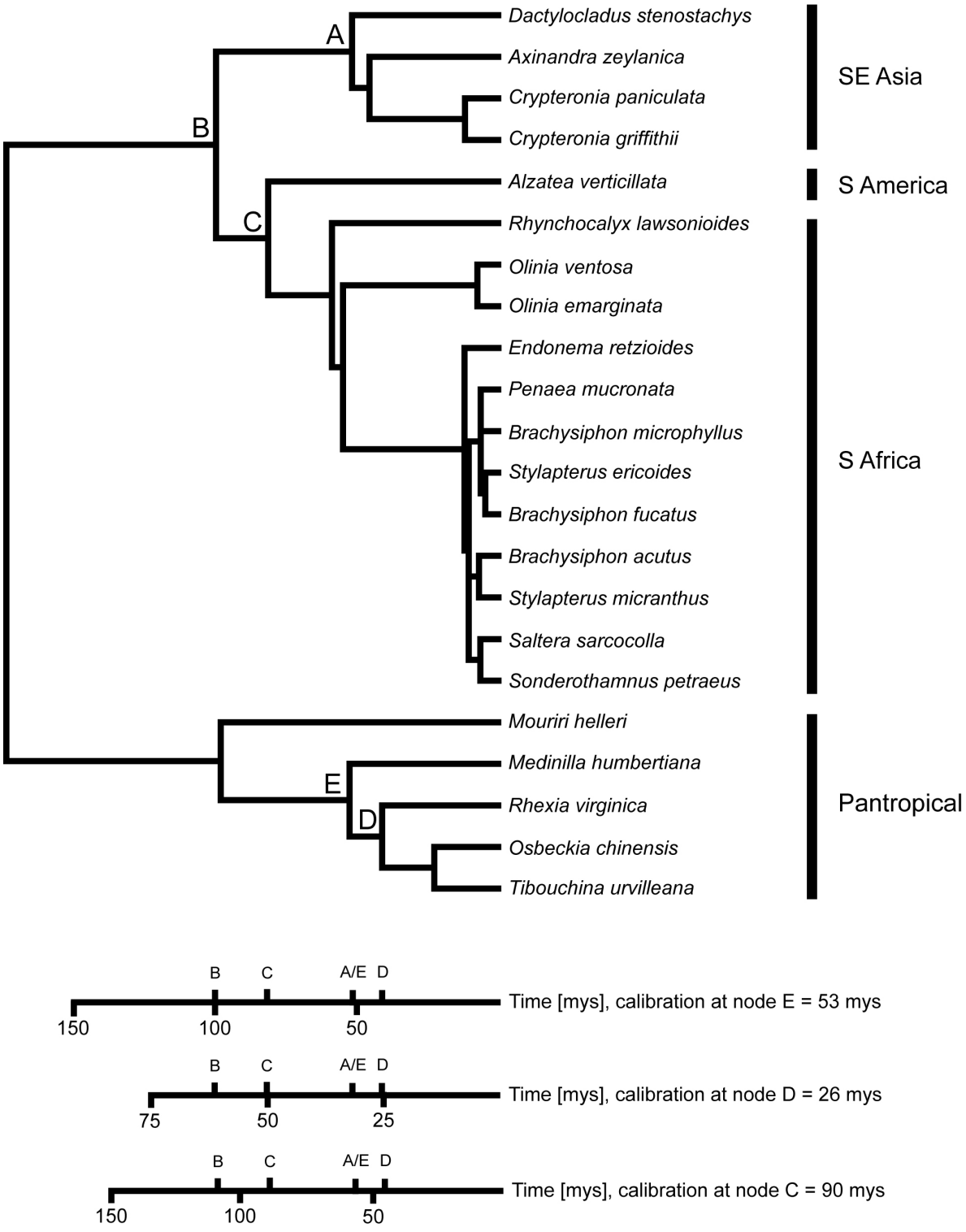
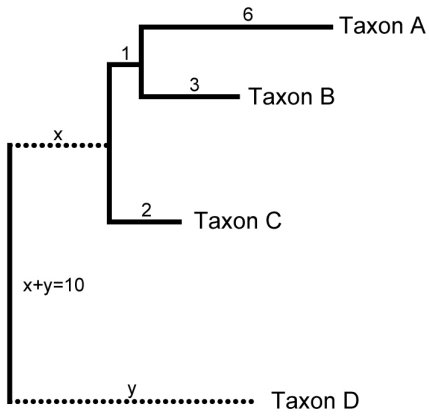
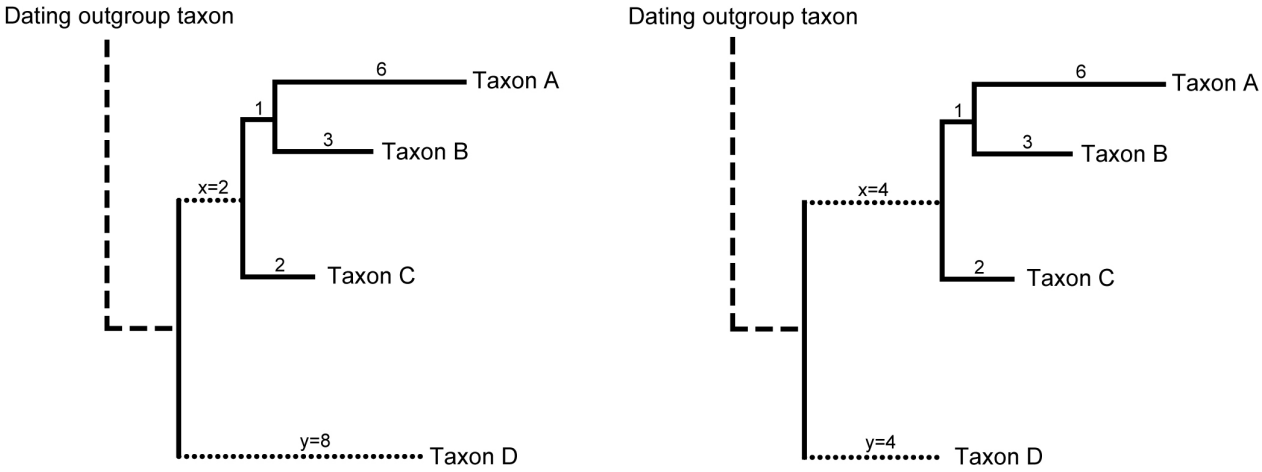


Figure 5b. Chronogram based on a Penalized Likelihood molecular dating analysis of the combined dataset. The independent use of three different calibration points resulted in three different time bars.



a) Additive tree to be analysed by molecular dating. Only the sum of the lenghts of the basal branches x and y is known, because the point where the root attaches to the basal branches is unknown.



b) An additional dating outgroup taxon was added to the tree to find where the root attaches to the basal branches. Two examples of possible root attachment points are shown, each determining different lenghts of the basal branches x and y. Note that the two trees differ also in the overall branch lengths from the root to the tips.

Figure 6. Use of a dating outgroup taxon to evaluate where the root attaches to the basal branches.

Chapter III.

Calibration of molecular clocks and the biogeographic history of Crypteroniaceae: A reply to Moyle

Elena Conti[1], Frank Rutschmann, Torsten Eriksson[2], Kenneth J. Sytsma[3], and David A. Baum[3]

[1]Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland

[2]Bergius Foundation, Royal Swedish Academy of Sciences, SE-104 05 Stockholm, Sweden

[3]Department of Botany, 430 Lincoln Drive,University of Wisconsin, Madison, WI 53706-1381, USA

Published in

Evolution (58): 1874-1876. 2004.

Reply to Robert G. Moyle

To test the molecular dating results and biogeographic interpretations reported by Conti *et al.* (2002), Moyle reanalyzed our published data set of 13 *rbc*L sequences representing five small taxa (the SE Asian Crypteroniaceae: the C clade; their W Gondwanan sister clade, formed by the S American *Alzatea* and the African *Rhynchocalyx*, Oliniaceae, and Penaeaceae: the AROP clade) and Melastomataceae. Using a single calibration point and nonparametric rate smoothing (NPRS; Sanderson, 2003), Moyle estimated an age of 68 mya (± 10.6 mys) for the split between Crypteroniaceae and the AROP clade, which contrasts with our published age of 116 mya (± 24 mys), obtained with fossil calibration and penalized likelihood (PL; Sanderson, 2003), and an age range of 50 to 151 mya, obtained by using three different calibration points and three different dating methods. Moyle concludes that his estimated age for the origin of the Crypteroniaceae stem lineage is "not congruent with a strict vicariance hypothesis for the distribution of Crypteroniaceae and nearest relatives" and that the differences in calibration "explain most of the differences in results".

While Moley's comment is timely by focusing on one of the most problematic issues in molecular dating analyses - namely, calibration -, we would like to highlight some weaknesses in his chosen analytical procedure, along with factual inaccuracies and misrepresentations of our original article. Furthermore, we report recent evidence from an expanded data set that includes *rbc*L, *ndh*F and *rpl*16 intron sequences from 22 species (henceforth, "expanded CAROP data set"; Rutschmann *et al.*, 2004) and offer some general reflections on the controversial issue of calibration in molecular dating analyses.

The criticism posed by Moyle that is most readily addressed concerns the phylogenetic placement of the South American *Alzatea*. Moyle's maximum likelihood (ML) analyses supported the placement of the S American *Alzatea* within the African clade, rather than as sister to the African clade as reported in Conti *et al.* (2002). He proposed that this discrepancy may be explained by the use of different models of nucleotide substitution in the two analyses, adding that low statistical support for the position of *Alzatea* suggests that its relationships remain unresolved. However, we would like to note that Moyle used outgroups (*Hauya*, Onagraceae and *Quisqualis*, Combretaceae) that are phylogenetically more distant from the CAROP/Melastomataceae clade than the outgroup we used (*Heteropyxis*, representing the sister clade of the CAROP/Melastomataceae clade; see Conti *et al.*, 1996). After re-running ML analyses with different combinations of outgroups and substitution models, we have concluded that instability in the position of *Alzatea* is caused primarily by different outgroup choices. In

support of our original result, recent ML analyses of the expanded CAROP data set corroborated the position of *Alzatea* as sister to the African clade with a bootstrap (BS) value of 86% (Rutschmann *et al.*, 2004). Furthermore, preliminary results based on combined nuclear 18S and 26S sequences of 9 taxa also confirmed the sister relationship of *Alzatea* to the African clade (Rutschmann and Conti, unpublished results).

A second point of disagreement concerns the properties and inferential value of the estimated age ranges. Moyle states: "Because of the wide range of age estimates produced by the different calibration points and molecular dating procedures, I re-examined the biogeographic history of Crypteroniaceae with particular attention to calibration procedure". He then elaborates on results based on a single dating method (NPRS) and a single calibration point (an age of 23 mya assigned to node E). This methodological approach will tend to provide a narrower range of estimated ages than would be obtained by using a range of methods, but such a superficially precise result may not by indicative of increased accuracy. Indeed, from a strictly analytical perspective, the choice of NPRS as the single dating method is questionable, since NPRS tends to overfit the data, especially when rates of molecular evolution change abruptly (Sanderson, 2002).

A more critical issue is, however, the way in which fossils are used to calibrate trees. Moyle used a single calibration point based on seeds that are dated at 23-26 mya and characterized by the large testa tubercles arranged in rows. These seeds have been assigned confidently to the crown group of Melastomeae, which were monophyletic in recent analyses (Renner and Meyer, 2001; Renner *et al.*, 2001; see also Collinson and Pingen, 1992). Yet, most likely due to scarce sampling, the three representatives of Melastomeae (*Tibouchina, Osbeckia*, and *Rhexia*) included in our *rbcL* analysis did not form a monophyletic group, but were members of a clade that also included *Medinilla* from the Dissochaeteae/ Sonerileae (Conti *et al.*, 2002, Figure 3). It was for this reason that we refrained from using this fossil as a calibration point. Moyle (see his Figure 1), in contrast, used the age of 23 ma to constrain node E, which subtends a clade formed by members of Melastomeae and Dissochaeteae/ Sonerileae. Since this node necessarily predates the origin of the Melastomeae stem group, to which the fossil seeds may be assigned (see also Renner and Meyer, 2001, Figure 3; Renner *et al.*, 2001, Figure 1), this improper calibration procedure automatically produces an underestimation of all nodal ages.

A further source of disagreement between Moyle's and our analyses concerns the nodal assignment of fossil leaves (dated at 53 mya) that are characterized by acrodromous venation, a synapomorphy exclusive for Melastomataceae among the sampled taxa (Renner *et al.*, 2001). Moyle criticizes our decision to use these fossil leaves to constrain the base of the Melastomataceae crown group (corresponding to node E in his Figure 1), instead of the base of its stem lineage (corresponding to node D in his Figure 1). Although we agree with Moyle that

the inclusion of basal lineages (e.g., *Pternandra*) would have been desirable (hence we are including them in on-going analyses), we disagree with his conclusion that these fossil leaves should be assigned to the base of the Melastomataceae stem lineage. The fossil leaves most closely resemble the leaves of extant Miconieae and Merianeae (Hickey, 1977), suggesting that they might be better assigned to shallower nodes within Melastomataceae (see also Rutschmann *et al.*, 2004. This observation, coupled with the fact that acrodromous venation is shared by all Melastomataceae, while their Memecylaceae sister group is characterized by brochidodromous venation, led us to conclude that, in the absence of additional information, the fossil leaves are more reasonably used to provide a minimal age for the crown rather than stem node of Melastomataceae (as done also by Renner, 2001 and Renner, 2004).

Some of the temporal uncertainties inherent to nodal assignment of fossils in a molecular phylogeny stem from the fact that the time of first appearance of distinctive synapomorphies in the fossil record postdates the origin of the group to which the fossils are assigned. Thus, when those fossils are used to provide minimal ages of the subtending stem lineages, and if those minimal ages are interpreted as estimates of actual ages, then one inevitably obtains a systematic underestimation of divergence times. The extent of the temporal gap between time of first appearance in the fossil record and origin of a group depends on several factors, including the fossil's probability of preservation, which in turn is influenced by properties inherent to the fossilized structures, the changing abundance through time of the structures being preserved in the fossil record, taphonomic idiosyncrasies (Morley and Dick, 2003), and the geologic characteristics of the stratigraphic layer where the fossil is retrieved (for a comprehensive and detailed review of calibration problems in molecular dating we refer the reader to Magallón, 2004; see also Graur and Martin, 2004). Ideally, it would be possible to estimate the difference between the observed age of the fossil and the "real" age of the group. Recently developed methods that attempt to achieve this goal make use of multiple lines of evidence, including the density and distribution of gaps in the fossil record, the number of extant species, the mean species lifetime, and clade diversification models. To our knowledge, these methods have been applied primarily to mammals (Tavaré *et al.*, 2002, Foote *et al.*, 1999).

In our original paper we also explored the results of assuming a correspondence between certain nodes and well-dated geological events, specifically equating the phylogenetic split between the S American *Alzatea* and its African sister clade with the formation of the South Atlantic. Moyle criticizes our use of geologic calibration as an example of circular reasoning. It is true that we did use geologic calibration to constrain this node, but always in conjunction with fossil calibrations (Conti *et al.*, 2002; Rutschmann *et al.*, 2004), subscribing to the practice of using as many calibration points as possible and then comparing and discussing the results

(Sanderson and Doyle, 2001; Thorne and Kishino, 2002; Yang and Yoder, 2003; Graur and Martin, 2004; Magallon, 2004). Furthermore, given his criticism, it is ironic that Moyle claimed support for his estimated age of the Crypteroniaceae stem lineage by noting its correspondence with the 68 mya age inferred (*sic!*) by Morley and Dick (2003). However, the latter authors did not estimate that divergence at 68 mya, but used it as a geologic constraint, marking the separation of India from Madagascar (see Figure 1 in Morley and Dick, 2003).

A further point of contention concerns how age ranges are used to reconstruct possible biogeographic scenarios. For example, Moyle criticizes the choice of using only the older portion (106-141 mya) of the inferred 50-151 mya range for our biogegraphic deductions (Conti *et al.*, 2002). However, the 106-141 range corresponded to the ages estimated by a variable-rate method (PL), while the younger portion of the age range included the ages estimated by two constant-rate methods (ML with clock enforced and Langley-Fitch; Sanderson, 2003). Because the likelihood ratio test (LRT) had rejected rate constancy, it seemed dubious to use clock-based age ranges for our biogeographic inferences.

Irrespective of the differences produced by different dating and calibration procedures, we would like to highlight some of Moyle's misrepresentations of our biogeographic conclusions. First, Moyle fails to explain the general context of our analyses: we used molecular dating estimates to test competing hypotheses on the Laurasian (Raven and Axelrod, 1974) vs. Gondwanan (Tobe and Raven, 1984) origin of Crypteroniaceae. Our results supported a Gondwanan origin for the family, a conclusion to which Moyle also subscribes. However, despite the apparent agreement between our general biogeographic conclusions, Moyle states: "The phylogeny and divergence time estimates produced here are not congruent with a strict Gondwanan vicariance hypothesis for the distribution of Crypteroniaceae and its nearest relatives. […] Instead, assuming a Gondwanan origin of the stem lineage leading to Crypteroniaceae and its allies, these dates infer (*sic!*) dispersal between Africa and India as India drifted northward". We were surprised to see that Moyle proposed this conclusion as a novel interpretation, in opposition to our purportedly strict vicariant explanation. Suffice it here to report the following summary statement: **"**Therefore, it seems reasonable to propose that Gondwanan drift played an essential role in the biogeographic history of Crypteroniaceae and related families and specifically that India, in its northward movement from Gondwana to Asia along the African coast, served as the most likely migration route for Crypteroniaceae" (p. 1940, Conti *et al.*, 2002).

In our more recent paper based on the expanded CAROP data set, we further developed the concept that dispersal of the Crypteroniaceae stem lineage from Africa to India might have been facilitated by the reduction of the ocean barrier between these two plates up to the Maastrichtian (Rutschmann *et al.*, 2004). Therefore, our interpretation of the biogeographic

history of Crypteroniaceae conforms to a geodispersalist (sensu Liebermann 1997, 2000), rather than a strictly vicariant model. While we remain open to the possibility that the use of different calibration points might produce relatively younger age estimates for the Crypteroniaceae stem lineage, we also note that Moyle's and our results both confirm a Cretaceous origin for this family and India's crucial role in its biogeographic history, as stated in the title of our original paper.

Moyle concludes that the main differences between his and our age estimates stem essentially from differences in calibration, seemingly implying that his calibration is right and ours is wrong. However, the complex issues revolving around nodal assignment of fossils cannot be easily reduced to a Manichaean view of scientific inference that relies on fixed categories of "right" and "wrong". The only kind of paleobotanical record that would unquestionably support the Gondwanan origin of Crypteroniaceae would be the retrieval of pre-Tertiary fossils attributable to Crypteroniaceae from Africa, Madagascar, or India. Barring that, we believe that biogeographic deductions should be based on multiple lines of evidence, drawn from both phylogenetic patterns and molecular dating, in combination with paleogeologic and paleoclimatic reconstructions and the evaluation of the potential for long-distance dispersal of the propagules We pursued this integrative, multi-faceted approach in our original and subsequent papers.

References

Conti, E., T. Eriksson, J. Schönenberger, K. J. Sytsma, and D. A. Baum. 2002. Early Tertiary out-of-India dispersal of Crypteroniaceae: evidence from phylogeny and molecular dating. *Evolution* **56**:1931-1942.

Conti, E., A. Litt, and K. J. Sytsma. 1996. Circumscription of Myrtales and their relationsships to other Rosids: evidence from *rbc*L sequence data. *American Journal of Botany* **83**:221-233.

Foote, M., J. P. Hunter, C. M. Janis, and J. J. Sepkoski Jr. 1999. Evolutionary and preservational constraints on origins of biologic groups: divergence times of Eutherian mammals. *Science* **283**:1310-1314.

Graur, D., and W. Martin. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *TRENDS in Genetics* **20**:80-86.

Hickey, L. J. 1977. Stratigraphy and paleobotany of the Golden Valley formation (Early Tertiary) of western North Dakota. Memoir 150. Geological Society of America, Boulder, Colorado.

Lieberman, B. S. 1997. Early Cambrian paleogeography and tectonic history: a biogeographic approach. *Geology* **25**:1039-1042.

Lieberman, B. S. 2000. Paleobiogeography. Using fossils to study global change, plate tectonics, and evolution. Kluwer Academic Publishers, New York.

Magallón, S. A. 2004. Dating lineages: molecular and paleontological approaches to the temporal framework of clades. *International Journal of Plant Sciences* **165**:7-21.

Morley, R. J., and C. W. Dick. 2003. Missing fossils, molecular clocks and the origin of the Melastomataceae. *American Journal of Botany* **90**:1638-1645.

Moyle, R. G. 2004. Calibration of molecular clocks and the biogeographic history of Crypteroniaceae. *Evolution* **58**:1871-1873.

Raven, P. H., and D. I. Axelrod. 1974. Angiosperm biogeography and past continental movements. *Annals of the Missouri Botanical Garden* **61**:539-673.

Renner, S. S. 2004. Bayesian analysis of combined chloroplast loci, using multiple calibrations, supports the recent arrival of Melastomataceae in Africa and Madagascar. *American Journal of Botany* **91**:1427-1435.

Renner, S. S., G. Clausing, and K. Meyer. 2001. Historical biogeography of Melastomataceae: the roles of Tertiary migration and long-distance dispersal. *American Journal of Botany* **88**:1290-1300.

Renner, S. S., and K. Meyer. 2001. Melastomeae come full circle: biogeographic reconstruction and molecular clock dating. *Evolution* **55**:1315-1324.

Ricklefs, R. E., and S. S. Renner. 1994. Species richness within families of flowering plants. *Evolution* **48**:1619-1636.

Rutschmann, F., T. Eriksson, J. Schönenberger, and E. Conti. 2004. Did Crypteroniaceae really disperse out of India? Molecular dating evidence from *rbc*L, *ndh*F, and *rpl*16 sequences. *International Journal of Plant Sciences* **165**:69-83.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a Penalized Likelihood approach. *Molecular Biology and Evolution* **19**:101-109.

Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**:301-302.

Sanderson, M. J., and J. A. Doyle. 2001. Sources of error and confidence intervals in estimating the age of Angiosperms from *rbc*L and 18S rDNA data. *American Journal of Botany* **88**:1499-1516.

Sytsma, K. J., A. Litt, M. L. Zjhra, J. C. Pires, M. Nepokroeff, E. Conti, J. Walker, and P. G. Wilson. 2004. Clades, clocks, and continents: historical and biogeographical analysis of Myrtaceae, Vochysiaceae, and relatives in the southern hemisphere. *International Journal of Plant Sciences* **165**:85-105.

Tavaré, S., C. R. Marshall, O. Will, C. Soligo, and R. D. Martin. 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature* **416**:726-729.

Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**:689-702.

Tobe, H., and P. H. Raven. 1984. The embryology and relationships of Oliniaceae. *Plant Systematics and Evolution* **146**:105-116.

Wikström, N., V. Savolainen, and M. W. Chase. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society B: Biological Sciences* **268**:2211-2220.

Yang, Z., and A. D. Yoder. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic Biology* **52**:705-716.

Chapter IV.
Tempo and mode of diversification in the African Penaeaceae and related taxa

Frank Rutschmann, Jürg Schönenberger[1], Peter Linder[2], and Elena Conti[2]

[1]*Department of Botany, Stockholm University, Lilla Frescativägen 5, SE-106 91 Stockholm, Sweden*

[2]*Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland*

In preparation for publication in

American Journal of Botany

Abstract

The Cape Floristic Region (CFR) in South Africa is characterized by very high levels of plant species diversity, including 69% endemics. Among these are the Penaeaceae (Myrtales), a small family of eriocoid shrubs most closely related to Oliniaceae and Rhynchocalycaceae, which also occur in the Cape, but further extend to Southern and Eastern Africa. Due to their isolated phylogenetic position and their narrow distribution, Penaeaceae have been referred to as ancient, paleoendemic elements. Previous phylogenetic studies revealed a pattern characterized by a long branch for the stem lineage, and many short branches for the crown group of Penaeaceae.

Here we present the results of phylogenetic and molecular dating analyses of eight chloroplast and three nuclear sequences aimed at reconstructing the tempo and mode of diversification in Penaeaceae and related families. Bayesian inference of divergence times and statistical evaluation of lineages through time plots provided a temporal framework for diversification. Maximum likelihood reconstruction of ancestral character states for leaf size and vegetation types offered new insights into possible evolutionary scenarios for their diversification.

The chloroplast and nuclear datasets produced significantly incongruent tree topologies, possibly suggesting earlier hybridization events followed by chloroplast capture. Molecular dating and the lineages through time analyses based on the plastid dataset showed that the crown radiation of Penaeaceae started about 17 mys ago, following the initiation of a climatic trend towards the modern seasonally cold and arid conditions in the Cape. Lignotubers and protective scleromorphic leaves possibly represent key innovations that triggered the radiation of Penaeaceae in the Middle Miocene, allowing Penaeaceae to grow on low nutrient soils. Contrary to Penaeaceae, Oliniaceae and Rhynchocalycaceae remained restricted to a temperate climate, and probably experienced a high degree of extinction. Our ancestral character state reconstructions indicate general shifts from tropical to fynbos vegetation types, and from macro- to leptophyllous leaves, corroborating the hypothesis that modern species, adapted to low nutrient soils and summer aridity, replaced an ancestral tropical flora during the Late Miocene aridification.

Key words: Diversification, radiation, Penaeaceae, Oliniaceae, Rhynchocalycaceae, biogeography, Cape Floristic Region, South Africa.

Introduction

The Cape Floristic Region (CFR) in South Africa is a biodiversity hotspot with more than 9000 plant species in an area of 90000 km^2. The level of endemism is very high: 68.8% of the species are restricted exclusively to the south-western tip of southern Africa (Goldblatt and Manning, 2000, 2002). The composition of this flora is very distinctive, with some families, for ex., Penaeaceae, Stilbaceae, Grubbiaceae, Roridulaceae, and Geissolomataceae found exclusively here. Other families, for ex., Iridaceae, Aizoaceae, Ericaceae, Proteaceae, and Restionaceae are more numerous and ecologically dominant in the CFR than in any other region of the world. Species richness is concentrated in a few large clades, which radiated mostly within the CFR (Linder, 2003; Linder and Hardy, 2004); 50 percent of the flora is contained in the 30 largest clades (Linder, 2003), and the flora is dominated by a few large genera (e.g. *Erica* with its 658 species; Goldblatt and Manning, 2000). The most characteristic vegetation type in the CFR is the fynbos, which occupies about half of the CFR and accounts for 80% of the species richness, including tall proteoids, heath-like ericoids, restioids, and bulbous geophytes (Goldblatt and Manning, 2002).

The dominating climate in the CFR is mediterranean, characterized by wet winters and warm, dry summers. This precipitation regime is more pronounced in the westernmost (ranging from the west coast to Swellendam) than in the eastern part of the CFR (ranging from the Overberg to Port Elizabeth; Schulze and McGee, 1978). In the western part, the winter weather is dominated by a succession of cold fronts and moist, nothwesterly winds, whereas in summer, high pressure cells and dry southeasterly winds dominate the climate. In the eastern part, seasons are less well defined in terms of winter rainfall and summer drought, because the southeasterly winds bring summer rain to the coastal plains and the mountains between the Overberg and the eastern Cape. In general, high altitude locations in the mountains are less exposed to seasonality and get more rain throughout the year than the lowlands (Goldblatt and Manning, 2002; Linder, 2003).

Most soils in the CFR are derived from quartzites and hard sandstones that form the Table Mountain and Witteberg groups and are generally deficient in nutrients (Kruger *et al.*, 1983). In the cool, high-rainfall mountain areas, rainfall leaches the scarce nutrients from the soil, leading to a bleached, extremely nutrient-poor horizon. At the base of the mountains, soils derived from Malmesbury and Bokkeveld shales are generally deeper and contain more clay and nutrients than the upper slopes and plateaux (Campbell, 1983; Goldblatt and Manning, 2000).

Because of its dry and high flammable nature, the fynbos is frequently exposed to bush fires, initiated by lightnings, rockfalls, or human activities. These hot and intense fires, proposed as a main driving force of diversification (Cowling, 1987; Van Wilgen, 1987; Manders and Cunliffe, 1987), cover large areas that burn down every 12 to 15 years on average, with extreme cycles of 5 to 50 years (in average every). Plants follow two main strategies to cope with fire: reseeding and resprouting (Pate *et al.*, 1989; Schutte *et al.*, 1995; Linder, 2003). Reseeders regenerate from seed and flower only after several years (Cowling, 1987, 1992), while resprouters usually regenerate from underground the roots or stems and flower within the first year or two after a fire. Both strategies are connected with complex and specific morphological adaptations that must have evolved under intense selective pressures (Linder, 2003).

Among the six families endemic to the CFR are the Penaeaceae (Myrtales), which comprise 23 species of shrubs and undershrubs in seven genera (Dahlgren, 1967a, 1967b, 1967c, 1968, 1971; Dahlgren and Thorne, 1984; Johnson and Briggs, 1984; Dahlgren and van Wyk, 1988; Schönenberger and Conti, 2003). Most of the species occur within a 200 km radius of Cape Town in the southwesternmost portion of the Cape Province (Carlquist and Debuhr, 1977) and inhabit the mountain fynbos, where they grow on rocky, oligotrophic sandstone rocks, slopes and cliffs under mesic conditions with more than 600 mm rainfall per year (Goldblatt and Manning, 2000; Schönenberger *et al.*, in press). A minor fraction of species occur along water streams (*Penaea dahlgrenii, Stylapterus ericoides, Stylapterus micranthus, Endonema laterifolia*) or in sandy flats (*Stylapterus fruticulosus*). All species grow on sandstone, except the limestone restricted *Brachysiphon mundii*. Due to their isolated taxonomic position among other plant groups endemic to the CFR and their small distribution range, Penaeaceae have been viewed as a paleoendemic floristic element, together with other families "with low evolutionary specialization" (Goldblatt and Manning, 2000) like Bruniaceae, Geissolomaceae, Grubbiaceae, Retziaceae, Roridulaceae, and Stilbaceae (Goldblatt and Manning, 2000). These paleoendemics are interpreted as relicts of an ancient temperate flora that once dominated the CFR and later progressively adapted to nutrient poor soils (see Discussion; Goldblatt and Manning, 2000).

Most closely related to Penaeaceae are the Southeast African families Oliniaceae and Rhynchocalycacae, both with fewer species, but a broader distribution range. Oliniaceae comprise about eight species in the single genus *Olinia* (Dahlgren and Thorne, 1984; Tobe and Raven, 1984b; Sebola and Balkwill, 1999; Conti *et al.*, 1996, 1997, 2002; Schönenberger and Conti, 2003; Rutschmann *et al.*, 2004, submitted). These trees and shrubs occur in the temperate aftromontane and coastal forests of eastern and southern Africa (South Africa, Swaziland, Moçambique, Uganda, Kenya, Tanzania, Malawi, Rwanda, Ethiopia, Republic of the Congo, Zambia, and Angola), and also on the island of St. Helena, where *Olinia ventosa* was most

probably introduced from South Africa (Cufodontis, 1960; Schönenberger and Conti, 2003). A sound taxonomic treatment of the family is not yet available (Sebola and Balkwill, 1999). Rhynchocalycaceae comprise the single species *Rhynchocalyx lawsonioides*, a rare, evergreen tree endemic to the South African provinces of Eastern Cape and KwaZulu-Natal (Van Beusekom-Osinga and van Beusekom, 1975; Johnson and Briggs, 1984; Tobe and Raven, 1984; Schönenberger and Conti, 2003). *Rhynchocalyx* is a 3 to 10 m high tree and grows mainly along the margins of temperate forests and close to watercourses and rivers (Palmer and Pitman, 1972). According to earlier molecular phylogenetic studies, *Rhynchocalyx* is sister to a clade comprising the monophyletic Penaeaceae and Oliniaceae (Conti *et al.*, 2002; Schönenberger and Conti, 2003; Rutschmann *et al.*, 2004). However, depending on gene and taxon sampling and the method used for phylogenetic inference, *Rhynchocalyx* might be nested within Oliniaceae, which are sister to Penaeaceae (Rutschmann *et al.*, submitted).

In the molecular phylogenetic studies mentioned above, a long branch between the stem lineage and the crown group of both Penaeaceae and Oliniaceae was observed in the phylograms, representing a high number of nucleotide substitutions. In subsequent molecular dating analyses aimed at reconstructing the biogeographic history of the southeast Asian Crypteroniaceae, closely related to Penaeaceae, Oliniaceae, and *Rhynchocalyx,* the long stem branches of Penaeceae and Oliniaceae persisted in the chronograms, indicating long time periods between their origin and diversification (Conti *et al.*, 2002; Rutschmann *et al.*, 2004; Rutschmann *et al.*, submitted). In addition, the chronograms suggested that after the described long period of absent diversification, both crown groups started to diversify rapidly. However, neither the specific timing, nor the tempo and mode of these diversifications have been investigated.

Here, we analyze an expanded nucleotide data set, comprising DNA sequences from eight plastid and three nuclear regions, with a combination of phylogenetic and molecular dating methods, lineage through time plots, and ancestral state reconstructions to elucidate diversification patterns in Penaeceae and Oliniaceae. More specifically, we address the following questions: i) What are the phylogenetic relationships within Penaeaceae, based on evidence from both the plastid and nuclear genomes?; ii) When did the stem lineages of Penaeaceae and Oliniaceae, respectively, originate and when did their crown groups start to diversify? iii) Did species in Penaeaceae and Oliniaceae accumulate at a constant rate through time or not? Do tempo and mode of diversification differ between the two families? iv) How can the observed diversification patterns be best explained in the light of the known history of the Cape flora?

Materials and Methods

Taxon sampling and DNA extractions

Taxon sampling, identical to that described in Schönenberger and Conti (2003), included a total of 31 taxa (see Table 1). Almost full sampling was achieved for Penaeaceae (19 of 23 species plus 4 subspecies of *Penaea cneorum*), Oliniaceae (5 of about 8 species), Rhynchocalycaceae (1 of 1 species), and Alzateaceae (1 of 1 species). The four missing Penaeaceae, *Stylapterus barbatus*, *Stylapterus dubius*, *Stylapterus sulcatus*, and *Stylapterus candolleanus*, are all either very rare, occur only in restricted areas, or were collected only once or twice. Following Sebola and Balkwill (1999), the three missing Oliniaceae are *Olinia discolor*, *Olinia rochetiana,* and *Olinia micranth*a. Total DNA was extracted as described in Schönenberger and Conti (2003) and Rutschmann *et al.* (2004).

In vitro amplification, sequencing and alignment

A total of 31 new Internal Transcribed Spacer (ITS) sequences were generated for this study, comprising ITS1, the 5.8S rRNA gene and ITS2. Taxon names, voucher information, and GenBank accession numbers are listed in Table 1. For all remaining sequences, voucher information and accession numbers are listed in Schönenberger and Conti (2003), Rutschmann *et al.* (2004), and Rutschmann *et al.*, submitted. (see Table 1) .

Our sampling of genetic loci was expanded from published studies (Schönenberger and Conti, 2003; Rutschmann *et al.*, 2004; Rutschmann *et al.*, submitted) to include 11 gene regions: the plastid *rbc*L exon, *ndh*F exon, *rpl*16-intron, *rps*16-intron, *trn*S-*trn*G spacer, *atp*B-*rbc*L spacer, *mat*K exon, and *psb*A-*trn*H spacer, plus the nuclear ribosomal 18S, 26S, and the nuclear ITS gene sequences (comprising the ITS1, the 5.8S rRNA gene and ITS2). To generate ITS sequences, we used amplification and sequencing primers slightly modified from Baum *et al.* (1998) and White *et al.* (1990): ITSLeu: 5'-GTC CAC TGA ACC TTA TCA TTT AG-3', ITS4: 5'-TCC TTC CGC TTA TTG ATA TGC-3', ITS3B: 5′-GCA TCG ATG AAG AAC GTA GC-3', ITS2: 5'-GCT GCG TTC TTC ATC GAT GC-3'. All other primer sequences and the procedures for PCR amplification and sequencing are described in Rutschmann *et al.* (2004) and Rutschmann *et al.*, submitted.

The software Sequencher 4.5 (Gene Codes, Ann Arbor, MI, USA) was used to edit and assemble complementary strands. Base positions were individually double-checked for agreement between the complementary strands. The sequences were first aligned using Clustal X 1.83 (Thompson *et al.*, 1997) prior to adjusting the alignments by eye with the software MacClade

4.07 (Maddison and Maddison, 2000). The chloroplast and nuclear DNA datasets used for further phylogenetic analyses contained 31 taxa and a total of 9955 aligned positions, composed of eight chloroplast and three nuclear partitions). The chloroplast dataset comprised sequences from the *rbc*L (1144 chars), *ndh*F (800 chars) and *mat*K (730 chars) genes, and the non-coding *rpl*16-intron (492 chars), *rps*16-intron (975 chars), *trn*S-*trn*G spacer (843 chars), *atp*B-*rbc*L spacer (1011 chars), and *psb*A-*trn*H spacer (576 chars; see Table 1). The nuclear dataset contained sequences from the ribosomal 18S (1630 chars) and 26S (962 chars) genes, and the non-coding ITS (792 chars; see Table 1).

To investigate possible intra-genomic variation of ITS repeats (Álvarez and Wendel, 2003), the PCR products of the ITS region from *Crypteronia paniculata*, *Olinia ventosa*, *Brachysiphon mundii*, and *Penaea mucronata* were each cloned by using the TOPO TA Cloning® kit by following the manufacturer's instructions (Invitrogen BV, Groningen). After growing the colonies at 37 °C overnight on agar plates with Luria-Bertani (LB) medium and ampicillin, four to five recombinant clones per plate were randomly selected and the ITS insert of each clone was in-vitro amplified using the provided universal primers M13F and M13R. The amplification products were then sequenced by using the same ITS primers and cycle-sequencing conditions described above.

Phylogenetic analyses

The chloroplast and nuclear datasets were analyzed separately. Bayesian analyses for topology and branch length estimation relied on MrBayes version 3.12 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). MrAIC 1.4 (Nylander, 2005), a program that uses PHYML (Guindon and Gascuel, 2003) to evaluate 24 nucleotide substitution models, was employed to select the best models based on the corrected Akaike information criterion (AICc; Burnham and Anderson, 2002; see Table 2). Parameter values for the chosen optimal models were then estimated simultaneously for each partition during topology and branch length optimization in MrBayes.

Bayesian topology estimation used one cold and three incrementally heated Markov chain Monte Carlo (MCMC) chains run for $1.5 * 10^6$ cycles, with trees sampled every 100^{th} generation, each using a random tree as a starting point and a temperature parameter value of 0.2 (the default in MrBayes). For each data set, MCMC runs were repeated twice. The first 5000 trees were discarded as burn-in. The remaining trees were used to construct one Bayesian consensus tree with mean branch lengths (Figures 1a and 1b) and to calculate the posterior probabilities for each branch. Examination of the logarithmic likelihoods and the observed consistency between the runs suggested that the used burn-in periods were sufficiently long.

Bootstrap support values were calculated by using the Perl script BootPHYML 3.4 (Nylander, 2005). This program (slightly modified by the first author) first generates 1000 pseudo-replicates in SEQBOOT (part of Phylip 3.63; Felsenstein, 2004), then performs a maximum likelihood analysis in PHYML (Guindon and Gascuel, 2003) for each replicate under the model of evolution selected above, and finally computes a 50% majority rule consensus tree by using CONSENSE (also part of the Phylip package).

In addition to the chloroplast and nuclear datasets, a separate phylogenetic analysis was performed as described above for the ITS dataset only, including 19 clones from *Crypteronia paniculata*, *Olinia ventosa*, *Brachysiphon mundii*, and *Penaea mucronata,* to check for monophyly of the respective clones.

To test for significant differences between the best tree topologies derived from the chloroplast and the nuclear datasets, the Kishino Hasegawa (KH) paired-sites test, also known as Resampling Estimated Log Likelihood (RELL) incongruence test (Kishino and Hasegawa, 1989; Felsenstein, 2004), was used. For both the chloroplast and the nuclear dataset, 10^5 sampled trees obtained in the MrBayes analyses were used to construct the respective 80% majoritiy rule consensus trees in PAUP* (Swofford, 2001). Using the plastid sequence data, the branch lengths for the two consensus trees were re-estimated under ML using the model parameters estimated during the MrBayes analysis. The Kishino-Hasegawa test was then used to compare the log likelihoods of the trees and test the null hypothesis that the trees are not significantly different. The same procedure was repeated starting with the nuclear sequence data. Because the KH paired-sites test indicated that the two trees were significantly different (see Results), molecular dating and ancestral state reconstruction analyses were performed on the data derived from the maternally-inherited chloroplast genome (Corriveau and Coleman, 1988)

Calibration

To calibrate the tree, we relied on previous results of broader phylogenetic and dating analyses in Myrtales (Rutschmann *et al.*, submitted, Rutschmann *et al.*, 2004, Conti *et al.*, 2002). Most recently, the age of the phylogenetic split between the African Penaeaceae/Oliniaceae and the South American Alzateaceae was placed at between 85.54 and 54.99 mys, based on the 97.5^{th} and 2.5^{th} percentile of the posterior distribution in Rutschmann *et al.* (submitted). By using such a broad time range for constraining the calibration node (node *c*; Figures 1a and 2), we tried to minimize problems related to secondary calibration (Graur and Martin, 2004; Hedges and Kumar, 2004).

Divergence time estimation

By using the χ^2 molecular clock test of Langley and Fitch (1974) implemented in r8s (Sanderson, 2003), we tested for rate constancy within the entire tree (global test), or within selected clades (local test; see Table 3). This χ^2 clock test compares the observed branch lengths with the branch lengths predicted under a clock model.

We then applied Bayesian dating (Thorne *et al.*, 1998; Thorne and Kishino, 2002) to estimate divergence times. This method uses a probabilistic model to describe the change in evolutionary rate over time and a Markov chain Monte Carlo (MCMC) routine to derive the posterior distribution of rates and time. The procedure we followed is divided into three steps and programs which are described in more detail in the Bayesian dating step-by-step manual version 1.5 (Rutschmann, 2005). It was performed on a 3 GHz Pentium IV machine running Ubuntu Linux 6.04. The values of the prior distributions in the third *multidivtime* step (Kishino *et al.*, 2001; Thorne and Kishino, 2002) were specified in units of 10 mys, as suggested in the manual: RTTM = 8.554, RTTMSD = 2, RTRATE and RTRATESD = 0.02525 (nuclear dataset), 0.00549 (plastid dataset), and 0.00585 (combined dataset), BROWNMEAN and BROWNSD = 0.117, and BIGTIME = 10. We ran the Markov chain for at least 10^6 cycles and collected one sample every 100 cycles, without sampling the first 10^5 cycles (burn-in sector). We performed each analysis at least twice with different initial conditions (INSEED parameters) and checked the output sample files to assure convergence of the Markov chain by using the program Tracer 1.3 (Rambaut and Drummond, 2005), although we realize that it is not possible to say with certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution (Cowles and Carlin, 1996). In addition, divergence times were also estimated by using penalized likelihood (Sanderson, 2002) implemented in *r8s* (Sanderson, 2003).

Diversification rate analysis

By using the chronogram derived from the plastid dataset (Figure 2), the number of diverging lineages was plotted against divergence time (LTT plots; Figures 3 and 4). To provide a global picture of diversification rates, the LTT analysis was first performed for all taxa, excluding the four subspecies of *Penaea cneorum* (full dataset; Figure 3). Additionally, we focused exclusively on members of the Penaeaceae crown group (again without subspecies) to provide more resolution for detecting diversification rate changes within the family (reduced dataset; Figure 4). The number of lineages was plotted in a decimal scale (Figure 3a and 4a) or log-transformed to linearize a potential exponentiality (Figure 3b and 4b).

Both the Cramér-von Mises and the Anderson-Darling goodness-of-fit tests implemented in APE (Paradis *et al.*, 2004) – a package written in R for the R project of statistical computing (R Development Core Team, 2004) - were used to test the null hypothesis that the diversification rates are constant over the entire ultrametric tree (Table 4). Both methods compare the empirical density function of the observations (branching times derived from the chronogram) to the expected exponential cumulative density function under the assumption of constant diversification rates (Paradis, 1998).

In addition, survival analysis (Paradis, 1997) implemented in APE was used to detect shifts in net diversification rate (Table 4). The method compares the likelihood of the data (branching times) under three different survival models. Model A assumes a constant rate of diversification (which leads to an exponential increase in lineages over time). Model B assumes a monotonically changing rate of diversification (Weibull law), and includes the estimated parameter β, which indicates if the diversification rate increases (positive β value) or decreases (negative β value). Model C assumes constant diversification rates, but assumes one abrupt change in the diversification rate at a single breakpoint (Paradis *et al.*, 2004). This breakpoint has to be specified a priori as parameter γ. For the global and the Penaeaceae crown group analyses, we set γ to 16.6 mys and 8.7 mys, respectively, because these values represented the most probable breakpoints based on a visual inspection of the LTT plots (Figures 3 and 4). The comparison among methods was done using both likelihood ratio tests (LRT) and the Akaike information criterion (AIC; see Table 4; Paradis, 1997; Kadereit *et al.*, 2004).

Ancestral character state reconstructions

In order to reconstruct character states for selected nodes of interest, an expanded chronogram from Rutschmann *et al.* (submitted) was used which comprises six more taxa of Crypteroniaceae (Figure 5). Although this chronogram is based on fewer genes, it provides better taxon sampling within Crypteroniaceae and its topology is almost entirely congruent with that of the chronogram based on eight chloroplast genes (see Figure 2). All 37 taxa were assigned to one out of three vegetation-type classes (tropical, temperate, and fynbos) and one out of five leaf-size classes (lepto-, nano-, micro-, meso-, and macrophyll; Table 5). Leaf sizes were calculated as the area of an ellipse (leaf length x leaf width x pi / 4) and then assigned to the appropriate leaf size classes following Raunkiaer (1916, 1934). Leaf dimensions were measured on herbarium specimens (see vouchers in Table 1) and derived from published literature (Palmer and Pitman, 1972; Pereira, 1996). Ancestral character states were then estimated under Maximum Likelihood optimization in Mesquite 1.06 (build g97; Maddison and Maddison, 2005) by using the Mk1

model (Table 6 and Figure 5). Additionally, leaf sizes were coded as continuous characters and reconstructed by using ML optimization in ANCML (Schluter *et al.*, 1997; Table 6 and Figure 5).

Results

Phylogenetic analyses

The two datasets used for phylogenetic analyses contained 31 taxa and a total of 9955 aligned positions, composed of eight chloroplast (6571 characters) and three nuclear gene partitions (3384 characters; see Table 1).

The optimal models of evolution, selected out of 24 different models with the corrected Akaike information criterion (implemented in MrAIC 1.4; Nylander, 2005), are summarized in Table 2. For both the chloroplast and the nuclear datasets, the corresponding MrBayes majority rule consensus trees with posterior mean branch lengths, maximum likelihood bootstrap support values obtained with BootPHYML 3.4 (above branches), and Bayesian posterior probabilities (below branches) are shown in Figures 1a and 1b.

The tree topologies generated from the plastid and nuclear data partitions were significantly different. Based on the plastid dataset, the KH paired-site test test (Kishino and Hasegawa, 1989) produced a –ln L score of 15044.91 for the plastid topology and a –ln L score of 15417.92 for the nuclear topology, leading to rejection of the null hypothesis that the two trees are not significantly different ($p < 0.05$). Likewise, based on the nuclear dataset, the KH test resulted in a –ln L score of 8342.07 for the nuclear topology, and in a –ln L score of 8469.34 for the plastid topology, which also led to rejection of the null hypothesis ($p < 0.05$).

The phylogenetic analysis of ITS sequences, including those generated from cloned PCR products, supported the monophyly of the clones amplified from *Crypteronia paniculata,* from *Olinia ventosa,* from *Brachysiphon mundii* and from *Penaea mucronata* (five clones each; results not shown). The sequences of ITS clones differed from each other by a minimum of 2 nucleotides (< 1%) between the clones from *Brachysiphon mundii* to a maximum of 6 nucleotides (about 1%) between the clones from *Penaea mucronata*. These results suggest that, at least for the tested species, mutations giving origin to different ITS repeats occurred after speciation, thus not affecting phylogenetic inference.

In agreement with other phylogenetic studies (Conti *et al.*, 1996, 1997 and 2002; Schönenberger and Conti, 2003; Rutschmann *et al.*, 2004; Rutschmann *et al.*, submitted), Oliniaceae and Penaeaceae form two monophyletic sister groups, based on both the chloroplast and the nuclear datasets (Figures 1a and 1b). The phylogenetic position of *Rhynchocalyx lawsonioides,* however, differs among partitions: in the plastid tree, it is sister to Oliniaceae, whereas in the nuclear tree it is sister to the clade formed by Oliniaceae and Penaeaceae. However, the relationships of *R. lawsonioides* are weakly supported in both phylogenies (BS =

53% and 45%, respectively). Phylogenetic incongruence was also observed within Penaeaceae: *Penaea* is monophyletic in the nuclear tree, but not in the plastid tree, because here *Penaea dahlgrenii* is sister to *Stylapterus ericoides*. Also the phylogenetic position of *Glischrocolla formosa* is dependent on the partition used for the phylogenetic reconstruction: in the plastid tree, it is sister to *Brachysiphon rupestris*, whereas in the nuclear tree, it is sister to *Endonema*. In both trees, *Endonema* is monophyletic, and *Saltera* and *Sonderothamnus* group together. The phylogenetic relationships of *Brachysiphon* and *Stylapterus* is problematic, because the taxa are arranged in clades with low statistical support. However, based on the present data, neither genus appears to be monophyletic.

Divergence time estimates

The Langley and Fitch χ^2 test applied to the plastid phylogram (after excluding *Crypteronia paniculata*, which was used as dating outgroup), indicated significant departure from rate constancy (Table 3). Rate constancy was clearly rejected also when the test was applied to subclades of the tree.

The estimated mean ages, their standard deviations, and the 95% credibility intervals based on the plastid dataset are reported for selected nodes of interest in Table 6, and a chronogram is shown in Figure 2. The use of penalized likelihood (implemented in *r8s*; Sanderson, 2002, 2003) instead of Bayesian dating (implemented in multidivtime; Kishino *et al.*, 2001; Thorne and Kishino, 2002) resulted in similar age estimates (data not shown).

Diversification rate analysis

By using the chronogram derived from the plastid dataset (Figure 2), the number of lineages was plotted against time (LTT plots). The plots for the full dataset (without subspecies) are shown in Figure 3, whereas the LTT plots for the reduced dataset (Penaeaceae crown group only) are shown in Figure 4. The diamonds in all plots represent mean estimated ages, while the dots represent the extremes of the credibility intervals. In Figures 3b and 4b, the number of lineages are log-transformed.

The LTT plots show two distinctive phases: The first is a period without any diversification between 53.47 mys ago and 16.6 mys ago (Figure 3), also represented by the long branches between the stem and the crown nodes of Penaeaceae and Oliniaceae in the chronogram (Figure 2). The second phase (16.6 mys ago to present; Figures 3 and 4) is characterized by a high, but non-expontential accumulation of lineages within the crown group of Penaeaceae.

Accordingly to these observations, both the Cramér-von Mises (CM) and the Anderson-Darling (AD) goodness-of-fit tests implemented in APE (Paradis *et al.*, 2004) rejected overall

diversification rate constancy on a 95% significance level for both datasets (Figures 3 and 4; full dataset: $p_{CM} < 0.01$, $p_{AD} < 0.025$; reduced dataset: $p_{CM} < 0.01$, $p_{AD} < 0.048$). Also the survival analysis (Paradis, 1997) confirmed the expectations from visual inspection of the LTT plots: For the full dataset (Figure 3), model C was selected among three different diversification models, based on the lower AIC values (Table 4). This model assumes an abrupt change in diversification 16.6 mys ago, but otherwise implies constant rates (Paradis *et al.*, 2004). For the reduced dataset (Penaeaceae crown group; Figure 4), model B was favored, again based on the lower AIC values (Table 4). According to its estimated β parameter ($\beta = 2.456$), this model assumes that the diversification rate decreased monotonically over time. The results of the LRT were either not significant or not decisive (see Table 3).

In summary, these statistical results suggest that the net diversification rate for Penaeaceae and Oliniaceae is not constant, but changes abruptly at about 16.6 mys ago (Figure 3). After this breakpoint, the diversification rate of Penaeaceae is higher than before, but decreases gradually over time, leading to a non-exponential increase of lineages (Figure 4).

Ancestral character state reconstructions

For the discrete characters (vegetation type class and leaf size class), the states with the highest proportional likelihoods are shown in Table 6. P values of less than 0.95 indicate that also other ancestral character states might be probable. In the small pie charts associated to the nodes in Figure 5, the proportional likelihoods of all possible states are shown (left side: vegetation type, right side: leaf size). The ancestral character state reconstructions show that the most recent common ancestor (MRCA) of all Penaeaceae (node *e*) likely had leptophyllous leaves and belonged to a fynbos-type vegetation, whereas the first member of the Oliniaceae crown group (node *g*) was microphyllous and belonged to a temperate vegetation. The MRCA of all Crypteroniaceae (node *b*) most likely had mesophyllous leaves and occurred in a tropical habitat. For the other nodes, the results are not significant, but a clear tendency is visible over time from ancestral plants with large leaves that grow in tropical habitats to phylogenetically younger plant groups with smaller leaves that occur in more arid habitats. This conclusion is supported by the gradient of the reconstructed leaf sizes treated as a continuous character in ANCML (Schluter *et al.*, 1997; see Table 6 and Figure 5, right side): the reconstructed leaf sizes are large (over 7000 mm^2) for the root node and the MRCA of Crypteroniaceae (node *b*), but they decrease for the MRCA's of Alzateaceae (node *c*), Oliniaceae (node *g*) and Penaeaceae (node *e*). For example, the leaf size of the MRCA of Crypteroniaceae (node *b*) is reconstructed to be 20 times bigger than that of the Penaeaceae MRCA (node *e*).

Discussion

Phylogenetic relationships within Penaeaceae

In his taxonomic treatment of the Penaeaceae, Dahlgren (1967a, 1967b, 1967c, 1968, 1971) subdivided the family into seven genera. Subsequently, two additional species, *Penaea dahlgrenii* (Rourke and McDonald, 1989) and *Brachysiphon microphyllus* (Rourke, 1995), were described. A recent molecular phylogenetic study based on six chloroplast markers partially contradicted earlier generic circumscriptions (Schönenberger and Conti, 2003). Specifically, it found the genera *Brachysiphon, Penaea,* and *Stylapterus* to be non-monophyletic. These results were not entirely surprising, as earlier authors had already found generic delimitations to be problematic (e.g., Supprian 1894, Gilg 1894), and also Dahlgren (1967a, 1968) pointed out several species to be "morphologically transitional" between genera, particularly so between *Brachysiphon* and *Stylapterus*. In addition, Rourke (1995) stated that the two newly described species, *Penaea dahlgrenii* and *Brachysiphon microphyllus*, were not easily assigned to any of Dahlgren's genera, further emphasizing the lack of clear-cut morphological boundaries between genera (see also Schönenberger *et al.*, in press). The phylogenetic relationships supported by the present analysis of chloroplast data (Figure 1a), which added two markers (*rbc*L and *ndh*F) to the data set of Schönenberger and Conti (2003), are largely identical to those of the former study. Results are also similar with respect to the overall resolution and branch support, which are still low in some areas of the tree.

The relationships inferred from the nuclear data set (Figure 1b), however, differ significantly from the chloroplast topology, but are likewise not congruent with the generic circumscriptions. Nevertheless, at least several parts of the two topologies are congruent, and these are probably reliable inferences of the species tree (Figures 1a and 1b): 1) *Endonema* is monophyletic and sister to the rest of the family, or at least belongs to the earliest diverging lineage in the family; 2) *Glischrocolla* also belongs to one of the earliest diverging lineages, either in a clade with *Brachysiphon rupestris* (cp-DNA, strong support; Figure 1a) or as sister to *Endonema* (nc-DNA, weak support; Figure 1b); 3) *Saltera* groups with *Sonderothamnus* in both topologies, but it remains unclear whether *Sonderothamnus* is paraphyletic to *Saltera*; 4) the genus *Penaea* sensu Dahlgren (1971) is monophyletic and the later described *Penaea dahlgrenii* is closely related to it, either as direct sister group (nc-DNA, strong support; Figure 1b) or perhaps as sister group together with *Stylapterus ericifolius* (cp-DNA, weak support; Figure 1a); and 5) *Penaea cneorum* is closest to *Penaea mucronata*, but it remains unclear whether various subspecies of *P. cneorum* are paraphyletic to *P. mucronata.* The remaining relationships – mainly

concerning the genera *Brachysiphon* and *Stylapterus* – are either not well supported or differ considerably between the two topologies.

Among the biological processes that might have caused the observed incongruencies between the chloroplast and nuclear trees of Penaeaceae and related taxa are orthology/paralogy conflation (e.g. Vanderpoorten *et al.*, 2004), lineage sorting (e.g. Doyle *et al.*, 2004), and hybridization (e.g. Rieseberg *et al.*, 1996). Our targeted cloning experiments on four species (*Crypteronia paniculata*, *Olinia ventosa*, *Brachysiphon mundii,* and *Penaea mucronata*) showed that all sequenced ITS repeats from a species are monophyletic, suggesting that duplication events and mutations that gave origins to the different repeats followed speciation. Therefore, in our case, orthology/paralogy conflation and lineage sorting appear to be a less likely primary source of incongruence between the two trees.

Hybridization, however, has been shown to be a widespread mode of speciation in plants at both the homoploid and allopolyploid levels, with fundamental effects on the potential incongruence between the maternally-inherited chloroplast and the biparentally-inherited nuclear phylogenies (Soltis *et al.*, 1992; Soltis and Soltis, 1999; Álvarez and Wendel, 2003). For example, the most conspicuous conflict among our gene trees involves *Stylapterus ericoides, Stylapterus micranthus,* and *Brachysiphon acutus*: in the nuclear topology (Figure 1b), the two former species form a strongly supported clade, whereas in the chloroplast topology (Figure 1a), the two latter species are strongly supported as being sisters. From a morphological point of view, the first hypothesis makes doubtless more sense as *Stylapterus micranthus* and *Brachysiphon acutus* differ in several floral characters (for discussion see Schönenberger and Conti, 2003). Thus, morphological evidence, together with the observed conflict between nuclear and chloroplast data, supports an earlier formulated hypothesis that the seemingly close relationship of *Stylapterus micranthus* and *Brachysiphon acutus* in the chloroplast phylogeny might be the result of an earlier hybridization event that was followed by chloroplast capture (Schönenberger and Conti, 2003).

For the remaining incongruent species, there is no further evidence from morphological traits and/or distribution ranges available, except a few chromosome counts, which do not concern incongruent taxa (2n = 20 for *Penaea mucronata*, *Penaea cneorum*, *Brachysiphon rupestris*; 2n = 40 for *Saltera sarcocolla*; Dahlgren, 1968, 1971).

Phylogenetic relationships within Oliniaceae

The taxonomy of the monogeneric Oliniaceae is not yet settled and species delimitation and the exact number of species are unclear (Sebola and Balkwill, 1999; von Balthazar and Schönenberger, in press). The five species included in the present study are grouped into two

strongly supported clades, which are both supported by chloroplast and nuclear data (Figures 1a and b). The same two clades were also found and discussed by Schönenberger and Conti (2003). In addition, the nuclear data set strongly supports a sister group relationship between *Olinia ventosa* and *O. capensis*, two sympatric species that have been recognized earlier to be closely related also based on their morphology (Sebola and Balkwill, 1999).

Diversification of Penaeaceae, Oliniaceae and Rhynchocalycaceae

By integrating our results from molecular dating (timing of diversification; Table 6 and Figure 2), LTT analyses (tempo and mode of diversification; Figures 3 and 4), and ancestral character state reconstructions (possible reasons of diversification; Tables 5 and 6 and Figure 5), we attempt to discuss the diversification history of Penaeaceae in the following paragraph. Before we proceed with the interpretation of our data, we would like to emphasize that we tried to minimize two pitfalls of the used methods: First, the use of secondary calibration in molecular dating is problematic because gene and taxon sampling and the methods used in the secondary study often differ significantly from those applied in the study that produced the original calibration point (Hedges and Kumar, 2003; Graur and Martin, 2004; Hedges and Kumar, 2004). In addition, the errors associated with divergence time estimation in the original study are often ignored when a specific age estimate is adopted for secondary calibration (Graur and Martin, 2004). Because we used a calibration point from one of our recent studies (Rutschmann *et al.*, submitted), which is based on a comparable gene and taxon sampling and identical methods, and because we fully incorporate error estimates into calibration, we think that the application of secondary calibration is appropriate in the present case. Secondly, it is known that LTT analyses are entirely dependent on taxon sampling, and their use is not recommended below a sampling level of 80% (Barraclough and Nee, 2001; Hawkins, 2006). For example, the five terminals representing Oliniaceae (Figure 2) are insufficient to infer diversification information from LTT plots (Figures 3 and 4), because we sampled only five out of eigth species (62.5%; von Balthazar and Schönenberger, in press). For Penaeaceae, however, we sampled 19 of 23 known species, providing enough data for investigating the tempo and mode of diversification for this group.

Our LTT analyses (see Figures 3 and 4) revealed two distinct phases in the diversification of Oliniaceae and Penaeaceae: i) a period without any diversification between 54 and 17 mys ago, also visible in the chronogram (Figure 2) as long branches between the stem and the crown nodes of Penaeaceae and Oliniaceae, and ii) a subsequent phase with a high accumulation of lineages within the crown groups of Penaeaceae and Oliniaceae (17 mys ago to present; see Figures 3 and 4).

For the first period, which spans the entire Oligocene and Eocene, we can only speculate about the history of the taxa examined here. For all three lineages, we observe an absence of diversification, which can be explained as either lack of speciation or a high amount of extinction, both leading to a stasis in the accumulation of lineages. Paleoclimatic reconstructions, although based on scant evidence, assume a changing climate during this period (Zachos *et al.*, 2001; DeConto and Pollard, 2003). It is imaginable, but not supported by fossil evidence, that numerous lineages of once more species-rich Penaeaceae, Oliniaceae and Rhynchocalycaceae could have gone extinct during this period, because they could not adapt rapidly enough to the changing climate. Alternatively, speciation was probably hindered because the necessary habitat was not available at that time. Our character state reconstructions suggest that early Penaeaceae, Oliniaceae and Rhynchocalycaceae (node *d*; Table 6 and Figure 5) possessed microphyllous leaves and inhabited a temperate or tropical environment at the beginning of that period (Table 6 and Figure 5). Over time, Penaeaceae developed smaller leaves (node *e*; Table 6 and Figure 5), whereas the foliage of Oliniaceae and Rhynchocalycaceae remained microphyllous (nodes *f* and *g*; Table 6 and Figure 5).

Around 16.6 mys ago (± 3.6 mys; see Figure 2), the second episode began: we observe an abrupt change in diversification (Table 4 and Figure 3; model C) at the beginning of the Middle Miocene. In Penaeaceae, but not in Oliniaceae and Rhynchocalycaceae, the number of lineages increases, leading to the 19 (out of 21) extant species. However, the diversification rate is not constant throughout this phase: according to our survival analysis (Table 4 and Figure 4), it decreases significantly over time (model B). Both the triggers for the radiation and the reasons for the subsequent decrease of the diversification rate are difficult to identify, because the climatic history of that period is not well known at sufficiently detailed temporal and spatial scales and fossils are not available (Linder, 2003). A popular paleoclimatic scenario is the aridification of the Cape in the Middle Miocene, caused by deteriorating world climatic conditions (Zachos et al, 2001). Although some key elements of this scenario are debated (e.g., the timing of the uplift of the Benguela current; Siesser, 1980; Siesser and Dingle, 1980), a climatic trend towards the modern seasonally cool and arid conditions in the Cape near the Miocene-Pliocene boundary is commonly accepted (Levyns, 1964; Linder *et al.*, 1992; Goldblatt and Manning, 2000; Linder, 2003, 2005).

Was the radiation of Penaeaceae about 17 mys ago driven by key innovations related to the described aridification scenario? At least according to the character state reconstructions, Penaeaceae possesed already small, thick, amphistomatous leaves at this time (node *e*; Figure 5 and Table 6; Rury and Dickison, 1984; Dickie and Gasson, 1999). While thick leaves provide improved gas exchange efficiency under water stress, amphistomatous leaves are generally

associated with plants growing in dry conditions (Parkhurst, 1978). Additionally, some Penaeaceae (e.g., *Saltera sarcocolla*) possess lignotubers (root-stocks), subterrestrial storage organs that allow them to endure periods of low water and nutrient availability and to resprout after fire (Carlquist and Debuhr, 1977; Schönenberger *et al.*, in press). On the other hand, Penaeaceae do not have adaptations like succulence or bulbs to endure long drought periods, but are restricted to mesic habitats like higher mountains, where they receive higher amounts of precipitation than in lower regions, or they occur close to water streams (Rourke, 1995). In addition, their small, scleromorphic leaves (see Table 5) can also be viewed as being primarily protective (Turner, 1994) and related to ensuring long lifespans for species growing slowly on nutrient-poor soils, rather than an adaptation to summer drought (Stock *et al.*, 1992).

Among other possible selective forces that may have triggered the radiation of Penaeaceae is also the adaptation to the fire regime. Penaeaceae are able to resprout efficiently after a fire, and seed germination has been shown to be stimulated chemically by charred wood or smoke (Keeley and Bond, 1997). However, the post fire mortality of Penaeaceae was shown to be high in comparison to other shrubs like *Aulax*, *Leucadendron* (Proteaceae) or *Widdringtonia* (Cupressaceae; Le Maitre *et al.*, 1992). In addition, Penaeaceae do not represent typical reseeders, which are usually taller and invest more resources in seed production than in resprouting (Pate *et al.*, 1989; Schutte *et al.*, 1995). In general, their ability to survive fire does not seem to be very efficient, at least not as efficient as the systems some competitors have evolved. As the timing of the first occurence of bush fires in the Cape is unknown (Linder, 2003; but see also Cowling, 1987, 1992), it is difficult to know if adaptation to fire became important in the Middle Miocene, or before or even after the radiation of Penaeaceae.

It is worth mentioning that Penaeaceae were not the only group that started to diversify rapidly at this time (16.6 mys ± 3.6 mys; 95% credibility interval: 24.8 – 10.7 mys; see Table 6). Among other Cape floristic elements that radiated in the Early and Middle Miocene are *Moraea* (Iridaceae; beginning of radiation dated to 22.5 ± 2.15 mys ago; Goldblatt, 2002), *Pelargonium* (Geraniaceae; diversification of xerophytic clade 18.5 to 11.4 mys ago; Bakker *et al.*, 2005), *Phylica* (Rhamnaceae; 13 to 14 mys ago; Richardson, 2004, but see Richardson *et al.*, 2001), and *Indigofera* (Fabaceae; 13 ± 3.5 mys ago; Schrire *et al.*, 2003). For *Phylica* and *Indigofera*, the plant specific traits that could have promoted the radiations were not investigated, but for *Pelargonium*, it is suggested that stem and leaf succulence and tuber formation represent important adaptations to summer aridity, which might have promoted the diversification of this genus (Bakker *et al.*, 2005). Similarily, the evolution of herbaceous forms with underground corms in *Moraea* are seen as adapations to a seasonally dry climate and nutrient poor soils (Goldblatt, 2002).

Oliniaceae and Rhynchocalycaceae, although much more widely distributed than Penaeaceae, are restricted to their purportedly ancient habitat, the temperate forest. Bearing much bigger leaves, these trees, up to 21 m high, grow only along water streams or in mesic mountain forests where continuous water and nutrient availability is guaranteed (in the CFR, two species occur along the coast and some rivers: *Olinia ventosa* and *Olinia capensis*; Goldblatt and Manning, 2000). The ancestral character state reconstructions suggest that this clade possessed microphyllous leaves and favoured a temperate habitat since the origin of its stem lineage about 58 mys ago (node *d*; Table 6 and Figure 5), with the fixation of these states in the MRCA of the Oliniaceae crown group (node *g*; Table 6 and Figure 5). We do not know why these trees speciated slower than Penaeaceae, but speculate that it might be the lack of preferred habitat space and their longer life cycle which prevented them from radiating as fast as Penaeaceae.

Conclusions

The diversification of the Cape flora can be expressed as a gradual transformation from species-poor clades with relictual taxa restricted to forests and fire protected habitats along water streams to species-rich clades with taxa that are highly adapted to open, heathy, periodically burnt habitats (Linder, 2003, 2005). Driven by complex climate changes since the late Cretaceous, gradual adaptations to drought and fire gave a selective advantage to the Cape floristic clades over the ancestral tropical flora (Linder and Hardy, 2004; see also Levyns, 1964; Linder *et al.*, 1992; Goldblatt and Manning, 2000).

The evolutionary trends revealed in our ancestral character state reconstructions for Penaeaceae, Oliniaceae, and Rhynchocalycaceae correspond to this model: they indicate shifts from a tropical/temperate habitat to the fynbos vegetation type, and from macro- and mesophyllous leaves to thick nano- and leptophyllous leaves (see Table 6 and Figure 5). Although the reasons for the sudden increase in diversification of Penaeaceae in the Middle Miocene remain speculative, lignotubers and protective scleromorphic leaves possibly represent key innovations that triggered their radiation, allowing Penaeaceae to grow on low nutrient soils.

The timeframe of the Penaeaceae radiation corresponds well to insights we have from other Cape taxa such as *Moraea*, *Pelargonium*, *Phylica* and *Indigofera*. It will be fascinating to see in the near future how the integration of diversification data from additional Cape clades, together with further paleoclimatic evidence, will contribute to a global reconstruction of the history of the Cape Flora.

References

Álvarez, I., and J. F. Wendel. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* **29**:417-434.

Bakker, F. T., A. Culham, E. M. Marais, and M. Gibby. 2005. Nested radiation in Cape *Pelargonium*. Pp. 75-100. Plant species-level systematics: new perspectives on pattern & process. A. R. G. Gantner Verlag, Ruggell, Liechtenstein.

Barraclough, T. G., and S. Nee. 2001. Phylogenetics and speciation. *Trends in Ecology & Evolution* **16**:391-399.

Baum, D. A., R. L. Small, and J. F. Wendel. 1998. Biogeography and floral evolution of Baobabs (*Adansonia*, Bombacaceae) as inferred from multiple data sets. *Systematic Biology* **47**:181-207.

Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Mulitmodel Inference: A Practical Information-Theoretic Approach. 2nd edition. Springer, New York.

Campbell, B. M. 1983. Montane plant environments in the fynbos biome. *Bothalia* **14**:283-298.

Carlquist, S., and L. Debuhr. 1977. Wood anatomy of Penaeaceae (Myrtales): comparative, phylogenetic, and ecological implications. *Botanical Journal of the Linnean Society* **75**:211-227.

Conti, E., T. Eriksson, J. Schönenberger, K. J. Sytsma, and D. A. Baum. 2002. Early Tertiary out-of-India dispersal of Crypteroniaceae: evidence from phylogeny and molecular dating. *Evolution* **56**:1931-1942.

Conti, E., A. Litt, and K. J. Sytsma. 1996. Circumscription of Myrtales and their relationsships to other Rosids: evidence from *rbc*L sequence data. *American Journal of Botany* **83**:221-233.

Conti, E., A. Litt, P. G. Wilson, S. A. Graham, B. G. Briggs, L. A. S. Johnson, and K. J. Sytsma. 1997. Interfamilial relationships in Myrtales: molecular phylogeny and patterns of morphological Evolution. *Systematic Botany* **22**:629-647.

Corriveau, J. L., and A. W. Coleman. 1988. Rapid Screening Method to Detect Potential Biparental Inheritance of Plastid DNA and Results for over 200 Angiosperm Species. *American Journal of Botany* **75**:1443-1458.

Cowles, M. K., and B. P. Carlin. 1996. Markov chain Monte Carlo convergence diagnosis: A comparative review. *Journal of the American Statistical Association* **91**:883-904.

Cowling, R. M. 1987. Fire and its role in coexistence and speciation in Gondwanan shrublands. *South African Journal of Science* **83**:106-112.

Cowling, R. M. 1992. Flora and vegetation. Pp. 23-61 *in* R. M. Cowling, ed. The Ecology of Fynbos: Nutrients, Fire and Diversity. Oxford University Press, Cape Town, South Africa.

Cufodontis, G. 1960. Die identifizierung von *Tephea* Delile und andere die Oliniaceae betreffende Feststellungen. *Österreichische Botanische Zeitschrift* **107**:106-112.

Dahlgren, R. 1967a. Studies in Penaeaceae. Part I. Systematics and gross mophology of the genus *Stylapterus* A. Juss. *Opera Botanica* **15**:3-40.

Dahlgren, R. 1967b. Studies on Penaeaceae III. The genus *Glischrocolla*. *Botaniska Notiser* **120**:57-68.

Dahlgren, R. 1967c. Studies on Penaeaceae IV. The genus *Endonema*. *Botaniska Notiser* **120**:69-83.

Dahlgren, R. 1968. Studies on Penaeaceae. Part II. The genera *Brachysiphon*, *Sonderothamnus* and *Saltera*. *Opera Botanica* **18**:5-72.

Dahlgren, R. 1971. Studies on Penaeaceae. VI. The genus *Penaea* L. . *Opera Botanica* **29**:5-58.

Dahlgren, R., and R. F. Thorne. 1984. The order Myrtales: circumscription, variation, and relationships. *Annals of the Missouri Botanical Garden* **71**:633-699.

Dahlgren, R., and A. E. van Wyk. 1988. Structures and relationships of families endemic to or centered in Southern Africa. *Monographs in Systematic Botany from the Missouri Botanical Garden* **25**:1-94.

DeConto, R. M., and D. Pollard. 2003. Rapid Cenozoic glaciation of Antarctica induced by declining atmospheric CO_2. *Nature* **421**:245-249.

Dickie, J. B., and P. E. Gasson. 1999. Comparative leaf anatomy of the Penaeaceae and its ecological implications. *Botanical Journal of the Linnean Society* **131**:327-351.

Doyle, J. J., J. L. Doyle, J. T. Rauscher, and A. H. D. Brown. 2004. Diploid and polyploid reticulate evolution throughout the history of the perennial soybeans (*Glycine* subgenus Glycine). *New Phytologist* **161**:121-132.

Felsenstein, J. 2004a. Inferring phylogenies. Sinauer Associates, Inc., Sunderland, MA.

Felsenstein, J. 2004b. PHYLIP (Phylogeny Inference Package), version 3.63. Department of Genetics, University of Washington, Seattle, WA.

Gilg, E. 1894. Penaeaceae. *Natürliche Pflanzenfamilien III* **6**:208-213. W. Engelmann, Leipzig.

Goldblatt, P., and J. C. Manning. 2000. Cape Plants. A Conspectus of the Cape Flora of South Africa. National Botanical Institute of South Africa, Pretoria, South Africa.

Goldblatt, P., and J. C. Manning. 2002. Plant diversity of the Cape Region of southern Africa. *Annals of the Missouri Botanical Garden* **89**:281-302.

Goldblatt, P., V. Savolainen, O. Porteous, I. Sostaric, M. Powell, G. Reeves, J. C. Manning, T. G. Barraclough, and M. W. Chase. 2002. Radiation in the Cape flora and the phylogeny of peacock irises *Moraea* (Iridaceae) based on four plastid DNA regions. *Molecular Phylogenetics and Evolution* **25**:341-360.

Graur, D., and W. Martin. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *TRENDS in Genetics* **20**:80-86.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**:696-704.

Hawkins, J. A. 2006. Using phylogeny to investigate the origins of the Cape flora: the importance of taxonomic, gene and genome sampling strategies. *Diversity and Distributions* **12**:27-33.

Hedges, S. B., and S. Kumar. 2003. Genomic clocks and evolutionary timescales. *TRENDS in Genetics* **19**:200-206.

Hedges, S. B., and S. Kumar. 2004. Precision of molecular time estimates. *Trends in Genetics* **20**:242-247.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* **17**:754.

Johnson, L. A. S., and B. G. Briggs. 1984. Myrtales and Myrtaceae - a phylogenetic analysis. *Annals of the Missouri Botanical Garden* **71**:700-756.

Kadereit, J. W., E. M. Griebeler, and H. P. Comes. 2004. Quaternary diversification in European alpine plants: pattern and process. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **359**:265-274.

Keeley, J. E., and W. J. Bond. 1997. Convergent seed germination in South African fynbos and Californian chaparral. *Plant Ecology* **133**:153-167.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution* **170-179**:170-179.

Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* **18**:352-361.

Kruger, F. J., D. T. Mitchell, and J. U. M. Jarvis. 1983. Mediterranean-type Ecosystems: The Role of Nutrients. Springer, Berlin.

Langley, C. H., and W. Fitch. 1974. An estimation of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution* **3**:161-177.

Le Maitre, D. C., C. A. Jones, and G. G. Forsyth. 1992. Survival of eight woody sprouting species following an autumn fire in Swatboskloof, Cape Province, South Africa. *South African Journal of Botany* **58**:405-413.

Levyns, M. R. 1964. Presidential address, migrations and origin of the Cape flora. *Transactions of the Royal Society of South Africa* **37**:85-107.

Linder, H. P. 2003. The radiation of the Cape flora, southern Africa. *Biological Reviews of the Cambridge Philosophical Society* **78**:597-638.

Linder, H. P. 2005. Evolution of diversity: the Cape flora. *TRENDS in Plant Science* **10**:536-541.

Linder, H. P., and C. R. Hardy. 2004. Evolution of the species-rich Cape flora. *Philosophical Transactions of the Royal Society London B* **359**:1623-1632.

Linder, H. P., M. E. Meadows, and R. M. Cowling. 1992. History of the Cape flora *in* R. M. Cowling, ed. The Ecology of Fynbos: Nutrients, Fire and Diversity. Oxford University Press, Cape Town, South Africa.

Maddison, P. G., and D. R. Maddison. 2000. MacClade 4: analysis of phylogeny and character evolution. Sinauer, Sunderland, MA.

Maddison, W. P., and D. R. Maddison. 2005. Mesquite: a modular system for evolutionary analysis. Available at http://mesquiteproject.org.

Manders, P. T., and R. N. Cunliffe. 1987. Fynbos plant life histories, populations dynamics and species interactions in relation to fire: an overview. Pp. 15-23 *in* R. M. Cowling, C. D. le Maitre, B. McKenzie, R. P. Prys-Jones and B. W. van Wilgen, eds. Distrubance and Dynamics of Fynbos Biome Communities. CSIR, Pretoria, South Africa.

Nylander, J. 2005a. BootPHYML 3.4. School of Computational Science (SCS), Florida State University, Tallahassee, Florida.

Nylander, J. 2005b. MrAIC 1.4. School of Computational Science (SCS), Florida State University, Tallahassee, Florida.

Palmer, E., and N. Pitman. 1972. Trees of Southern Africa covering all known indigenous species in the Republic of South Africa, South-West Africa, Botswana, Lesotho and Swaziland. A. A. Balkema, Cape Town, South Africa.

Paradis, E. 1997. Assessing temporal variations in diversification rates from phylogenies: estimation and hypothesis testing. *Proceedings of the Royal Society of London B* **264**:1141-1147.

Paradis, E. 1998. Testing for constant diversification rates using molecular phylogenies: A general approach based on statistical tests for goodness of fit. *Molecular Biology and Evolution* **15**:476-479.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**:289-290.

Parkhurst, D. F. 1978. The adaptive significance of stomatal occurrence on one or both surfaces of leaves. *Journal of Ecology* **66**:367-383.

Pate, J. S., R. H. Froend, B. J. Bowen, A. Hansen, and J. Kuo. 1989. Seedling growth and storage characteristics of seeder and resprouter species of Mediterranean-type ecosystems of S.W. Australia. *Annals of Botany* **65**:585-601.

Pereira, J. T. 1996. Crypteroniaceae. Pp. 135-149 *in* E. Soepadmo, K. M. Wong and L. G. Saw, eds. Tree flora of Sabah and Sarawak, Volume 2. Forest Research Institute Malaysia, Sabah Forestry Department, and Sarawak Forestry Department, Kuala Lumpur, Malaysia.

R Development Core Team. 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org.

Rambaut, A., and A. J. Drummond. 2005. Tracer 1.3. A program for analyzing results from Bayesian MCMC programs such as BEAST & MrBayes. Department of Zoology, University of Oxford, Oxford, UK. Available from http://evolve.zoo.ox.ac.uk/software.html.

Raunkiaer, C. 1916. The use of leaf size in biological plant geography. Botanisk Tidsskrifit, 34. Pp. 368–378 *in* C. Raunkiaer, ed. Life forms of plants and statistical plant geography. Arno Press, New York.

Raunkiaer, C. 1934. Life forms of plants and statistical plant geography. Arno Press, New York.

Richardson, D. 2004. Evolution of *Phylica* (Rhamnaceae) in the Cape. *Abstract for the meeting "Recent Floristic Radiations in the Cape Flora", 3-5 July 2004, Zurich, Switzerland.*

Richardson, J. E., F. M. Weitz, M. F. Fay, Q. C. B. Cronk, H. P. Linder, G. Reeves, and M. W. Chase. 2001. Rapid and recent origin of species richness in the Cape flora of South Africa. *Nature* **412**:181-183.

Rieseberg, L. H., J. Whitton, and C. R. Linder. 1996. Molecular marker incongruence in plant hybrid zones and phylogenetic trees. *Acta Botanica Neerlandica* **45**:243-262.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**:1572-1574.

Rourke, J. P. 1995. A new species of Brachysiphon (Penaeaceae) from the Southern Cape, South Africa. *Nordic Journal of Botany* **15**:63-66.

Rourke, J. P., and D. J. McDonald. 1989. A new species of *Penaea* (Penaeaceae), from the Langeberg range, southern Cape. *South African Journal of Botany* **55**:400-404.

Rury, P. M., and W. C. Dickison. 1984. XI. Structural correlations among wood, leaves and plant habitat. Pp. 495-540 *in* R. A. White and W. C. Dickison, eds. Contemporary Problems in Plant Anatomy. Academic Press, London.

Rutschmann, F. Assessing calibration uncertainty in molecular dating: The assignment of fossils to alternative calibration points. *Submitted.*

Rutschmann, F. 2005. Bayesian molecular dating using paml/multidivtime. A step-by-step manual. Version 1.5 (July 2005). Institute of Systematic Botany, University of Zurich, Zurich, Switzerland. Available from http://www.plant.ch.

Rutschmann, F., T. Eriksson, J. Schönenberger, and E. Conti. 2004. Did Crypteroniaceae really disperse out of India? Molecular dating evidence from *rbc*L, *ndh*F, and *rpl*16 sequences. *International Journal of Plant Sciences* **165**:69-83.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a Penalized Likelihood approach. *Molecular Biology and Evolution* **19**:101-109.

Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**:301-302.

Schluter, D., T. Price, A. O. Mooers, and D. Ludwig. 1997. Likelihood of ancestor states in adaptive radiation. *Evolution* **51**:1699-1711.

Schönenberger, J., and E. Conti. 2003. Molecular phylogeny and floral evolution of Penaeaceae, Oliniaceae, Rhynchocalycaceae, and Alzateaceae (Myrtales). *American Journal of Botany* **90**:293-309.

Schönenberger, J., E. Conti, F. Rutschmann, and R. Dahlgren. *In press*. Penaeaceae *in* K. Kubitzki, ed. The families and genera of vascular plants. Vol. 9. Springer, Berlin.

Schrire, B. D., M. Lavin, N. P. Barker, H. Cortes-Burns, I. von Senger, and J.-H. Kim. 2003. Towards a phylogeny of *Indigofera* (Leguminosae-Papilionoideae): identification of major clades and relative ages. . Pp. 269-302 *in* B. B. Klitgaard and A. Bruneau, eds. Advances in Legume Systematics, part 10, Higher Level Systematics. Royal Botanic Gardens, Kew, London.

Schulze, R. E., and O. S. McGee. 1978. Climatic indices and classifications in relations to the biogeography of southern Africa. Pp. 19-52 *in* M. J. A. Werger, ed. Biogeography and Ecology of Southern Africa, Chapter 2. W. Junk, The Hague, Netherlands.

Schutte, A. L., J. H. J. Vlok, and B.-E. Van Wyk. 1995. Fire-survival strategy - a strategy of taxonomic, ecological and evolutionary importance in fynbos legumes. *Plant Systematics and Evolution* **195**:243-259.

Sebola, R. J., and K. Balkwill. 1999. Resurrection of two previously confused species, *Olinia capensis* (Jacq.) Klotzsch and *O. micrantha* Decne. (Oliniaceae). *South African Journal of Botany* **65**:97-103.

Siesser, W. G. 1980. Late Miocene origin of the Benguela upwelling system of northern Namibia. *Science* **208**:283-285.

Siesser, W. G., and R. V. Dingle. 1980. Tertiary sea-level movements around southern Africa. *Journal of Geology* **89**:83-96.

Soltis, D. E., and P. S. Soltis. 1999. Polyploidy: recurrent formation and genome evolution. *Trends in Ecology & Evolution* **14**:348-352.

Soltis, D. E., P. S. Soltis, and J. A. Doyle. 1992. Molecular data and polyploid evolution in plants. Pp. 177-201 *in* D. E. Soltis, P. S. Soltis and J. A. Doyle, eds. Molecular systematics of plants. Chapman & Hall, New York, USA.

Stock, W. D., F. van der Heyden, and O. A. M. Lewis. 1992. 9. Plant structure and function. Pp. 226-240 *in* R. M. Cowling, ed. The Ecology of Fynbos: Nutrients, Fire and Diversity. Oxford University Press, Cape Town, South Africa.

Supprian, K. 1894. Beiträge zur Kenntnis der Thymelaeaceae und Penaeaceae. *Botanisches Jahrbuch* **18**:306-353.

Swofford, D. L. 2001. PAUP* 4.0b10: Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer, Sunderland, MA.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **24**:4876-4882.

Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**:689-702.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**:1647-1657.

Tobe, H., and P. H. Raven. 1984. The embryology and relationships of Oliniaceae. *Plant Systematics and Evolution* **146**:105-116.

Turner, I. M. 1994. Sclerophylly: primarily protective? *Functional ecology* **8**:669-675.

van Beusekom-Osinga, R. J., and C. F. van Beusekom. 1975. Delimitation and subdivision of the Crypteroniaceae (Myrtales). *Blumea* **22**:255-266.

Van Wilgen, B. W. 1987. Fire regimes in the Fynbos biome. Pp. 6-14 *in* R. M. Cowling, C. D. le Maitre, B. McKenzie, R. P. Prys-Jones and B. W. van Wilgen, eds. Disturbance and Dynamics of Fynbos Biome Communities. CISR, Pretoria, South Africa.

Vanderpoorten, A., A. J. Shaw, and C. J. Cox. 2004. Evolution of multiple paralogous adenosine kinase genes in the moss genus *Hygroamblystegium*: phylogenetic implications. *Molecular Phylogenetics and Evolution* **31**:505-516.

von Balthazar, M., and J. Schönenberger. *In press.* Oliniaceae *in* K. Kubitzki, ed. The families and genera of vascular plants. Vol. 9. Springer, Berlin.

White, T. J., T. Birns, S. Lee, and J. Taylor. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. Pp. 315-322 *in* M. Innis, D. Gelfand, J. Sninsky and T. White, eds. PCR protocols: a guide to methods and applications. Academic Press, San Diego, CA, USA.

Zachos, J., M. Pagani, L. Sloan, E. Thomas, and K. Billups. 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**:686-693.

Tables

Table 1. Species names, sources and GenBank Accession numbers of newly amplified ITS DNA sequences. Herbaria acronyms: BOL = Bolus, University of Cape Town, South Africa; Z = University of Zurich, Switzerland. Sources of all other gene sequences used in this study are listed in Schönenberger and Conti (2003), Rutschmann *et al.* (2004), and Rutschmann *et al.*, submitted.

Taxon	Voucher	ITS GenBank accession numbers
Crypteroniaceae A. DC. (1868), nom. cons.		
Crypteronia paniculata Blume	M. W. Chase 1235, cultivated in Bogor Botanic Garden VIII.B.67	AM235848
Alzateaceae S.A. Graham (1985)		
Alzatea verticillata Ruiz & Pavon	J. Gomez-Laurito, s.n., (USJ)	AM235849
Rhynchocalycaceae L.A.S. Johnson & B.G. Briggs (1985)		
Rhynchocalyx lawsonioides Oliver	T. Abbott 7658, (Z)	AM235850
Oliniaceae Arn. ex Sond. (1839), nom. cons.		
Olinia capensis (Jacq.) Klotzsch	J. Schönenberger 519, (Z), (BOL)	AM235851
Olinia emarginata Burtt Davy	J. Schönenberger 579, cultivated in Kirstenbosch Botanical Garden, (Z)	AM235852
Olinia radiata Hofmeyr & Phill.	T. Abbott, 6341, (Z)	AM235853
Olinia vanguerioides Baker f.	A. Blarer, s.n., (Z)	AM235854
Olinia ventosa (L.) Cufod.	J. Schönenberger 378, (Z),(BOL)	AM235855
Penaeaceae Sweet ex. Guillemin (1828), nom. cons.		
Brachysiphon acutus (Thunb.) A. Juss.	J. Schönenberger 365, (Z), (BOL)	AM235856
Brachysiphon fucatus (L.) Gilg	J. Schönenberger 357, (Z), (BOL)	AM235857
Brachysiphon microphyllus Rourke	J. Schönenberger 386, (Z), (BOL)	AM235858

Brachysiphon mundii Sond.	J. Schönenberger 377, (Z), (BOL)	AM235859
Brachysiphon rupestris Sond.	J. Schönenberger 366, (Z), (BOL)	AM235860
Endonema lateriflora (L.f.) Gilg	J. Schönenberger 369, (Z), (BOL)	AM235861
Endonema retzioides Sond.	J. Schönenberger 370, (Z), (BOL)	AM235862
Glischrocolla formosa (Thunb.) R. Dahlgren	J. Schönenberger 521, photo-vouchered	AM235863
Penaea acutifolia A. Juss.	J. Schönenberger 376, (Z), (BOL)	AM235864
Penaea cneorum Meerb. ssp. *cneorum*	J. Schönenberger 363, (Z), (BOL)	AM235865
Penaea cneorum Meerb. ssp. *gigantea* R. Dahlgren	J. Schönenberger 375, (Z), (BOL)	AM235866
Penaea cneorum Meerb. cf. ssp. *lanceolata* R. Dahlgren	J. Schönenberger 320, (Z), (BOL)	AM235867
Penaea cneorum Meerb. ssp. *ovata* (Eckl. & Zeyh. ex A. DC.) R. Dahlgren	J. Schönenberger 374, (Z), (BOL)	AM235868
Penaea cneorum Meerb. cf. ssp. *ruscifolia* R. Dahlgren	J. Schönenberger 368, (Z), (BOL)	AM235869
Penaea dahlgrenii Rourke	J. Schönenberger 388, (Z), (BOL)	AM235870
Penaea mucronata L.	J. Schönenberger 354, (Z),(BOL)	AM235871
Saltera sarcocolla (L.) Bullock	J. Schönenberger 360, (Z), (BOL)	AM235872
Sonderothamnus petraeus (Barker f.) R. Dahlgren	J. Schönenberger 362, (Z), (BOL)	AM235873
Sonderothamnus speciosus R. Dahlgren	J. Schönenberger 364, (Z), (BOL)	AM235874
Stylapterus ericifolius (A. Juss.) R. Dahlgren	J. Schönenberger 372, (Z), (BOL)	AM235875
Stylapterus ericoides A. Juss. ssp. *pallidus* R. Dahlgren	J. Schönenberger 355, (Z), (BOL)	AM235876
Stylapterus fruticulosus (L. f.)	J. Schönenberger 359, (Z), (BOL)	AM235877
Stylapterus micranthus R. Dahlgren	M. Johns, s.n., (Z)	AM235878

Table 2. Gene partitions and models of molecular evolution selected for the plastid and the nuclear dataset by using MrAIC (Nylander, 2005) by the application of the corrected Akaike information criterion (AICc). The corresponding model parameters were estimated in MrBayes (Huelsenbeck and Ronquist, 2001) during the phylogenetic reconstruction.

Dataset	Gene partition	Size	Optimal model obtained by using AICc selection (in MrAIC)
Chloroplast	*rbc*L	1-1144 (1144 chars)	GTR+G
	*ndh*F	1145-1944 (800 chars)	GTR+G
	*rpl*16-intron	1945-2436 (492 chars)	HKY
	*rps*16-intron	2437-3411 (975 chars)	GTR+G
	*trn*S-*trn*G	3412-4254 (843 chars)	GTR+G
	*atp*B-*rbc*L spacer	4255-5265 (1011 chars)	GTR+G
	*mat*K	5266-5995 (730 chars)	GTR+G
	*psb*A-*trn*H	5996-6571 (576 chars)	GTR+G
Nuclear	nr18S	1-1630 (1630 chars)	SYM+I+G
	nr26S	1631-2592 (962 chars)	GTR+I+G
	ITS	2593-3384 (792 chars)	HKY+G

Table 3. Estimates of rate heterogeneity for the entire phylogram and selected clades, evaluated by using the χ^2 clock test of Langley and Fitch (1974) implemented in r8s (Sanderson, 2003). χ^2 represents the calculated test value, df the degrees of freedom (number of taxa – 1), and p the significance level. Results are based on the chloroplast dataset.

Clade	χ^2	df	p	Constant molecular clock
Entire phylogram	158.2	29	$9.75*10^{-20}$	rejected
Penaeaceae crown group	94.3	22	$6.3*10^{-11}$	rejected
Oliniaceae crown group	16.24	4	$2.71*10^{-3}$	rejected

Table 4. Results of the survival analysis (Paradis, 1997) implemented in APE (Paradis *et al.*, 2004). For the full dataset (see Figure 3), none of the three models is preferred by the likelihood ratio tests (LRT; p values > 0.05), but the Akaike information criterion (AIC) favours model C. For the reduced dataset (see Figure 4), either model B or C are preferred by the LRT's (p values < 0.05), whereas the AIC favors model B. The values being decisive for model selection are bold.

Model		**Full dataset (Figure 3)**		**Reduced dataset (Figure 4): Penaeaceae crown group only**	
		log likelihood	**AIC**	**log likelihood**	**AIC**
A	diversification rate is constant over time	-33.51	69.02	-16.12	34.23
B	diversification rate either increases ($\beta > 0$) or decreases ($\beta < 0$) monotonically	-33.49 $\beta = 0.97$ ($p = 0.845$)	70.98	-8.61 $\beta = 2.456$ ($p = 0.0001$)	**21.22**
C	diversification rate is constant, but changes abruptly at the given breakpoint γ	-31.71 $\gamma = 16.6$ mys ($p = 0.058$)	**67.42**	-20.62 $\gamma = 8.7$ mys ($p = 0.00269$)	45.23

Table 5. Habitat, distribution, maximal leaf size, assigned leaf size class and vegetation type class of the taxa used in this study for the reconstruction of ancestral character states (Table 6 and Figure 5). Distribution data is mainly from Pereira (1996); Palmer and Pitman (1972), and Goldblatt and Manning (2000). The phytogeographic centers of the Cape Floristic Region (CFR; Goldblatt and Manning, 2000) are abbreviated as follows: SW: Southwest, NW: Northwest, SE: Southeast, LB: Langeberg, KM: Karoo mountain. Leaf sizes were calculated as the area of an ellipse (leaf length x leaf width x pi / 4). Leaf size classes follow the ecological definition by Raunkiaer (1916, 1934).

Taxon	**Habitat**	**Distribution**	**Max. leaf size** (area in mm^2)	**Leaf size class**	**Vegetation type class**
Dactylocladus stenostachys	Peat swamp and mixed swamp forests	Borneo: Sabah, Sarawak, Brunei, Kalimantan	5497.79	mesophyll	tropical
Crypteronia paniculata	Tropical hill forests	Contintal SE Asia, Andaman Islands, Myanmar, Peninsular Malaysia, Borneo, Sumatra, Java, Lesser Sunda Islands, Philippines	16493.36	mesophyll	tropical
Crypteronia borneensis	Mixed dipterocarp forest	Borneo: Sabah, Sarawak and Brunei	40840.70	macrophyll	tropical
Crypteronia griffithii	Primary tropical forest	Myanmar, Central Sumatra, Peninsualr Malaysia, Borneo	47123.89	macrophyll	tropical
Crypteronia glabrifolia	Mixed dipterocarp forest,	Borneo: Sarawak and Brunei	12566.37	mesophyll	tropical
Axinandra zeylanica	Lower montaine tropical rainforest	Sri Lanka	3534.29	mesophyll	tropical
Axinandra coriacea	Lowland or lower montaine rainforest	Borneo: Sabah, Sarawak and Kalimantan	8246.68	mesophyll	tropical
Alzatea verticillata	Tropical mountain forest	Costa Rica, Panama, Equador, Peru, Bolivia	9424.78	mesophyll	tropical
Rhynchocalyx lawsonioides	Margins of temperate forests, close to watercourses	Eastern Cape and KwaZulu-Natal	1413.72	microphyll	temperate
Olinia ventosa	Coastal forests	South Africa: Western and Eastern	2324.78	mesophyll	temperate

		Cape Provinces, St. Helena (introduced)			
Olinia capensis	Coastal forests	South Africa: Western and Eastern Cape Provinces	954.26	microphyll	temperate
Olinia radiata	Mountain forests, wooden kloofs	South Africa: KwaZulu-Natal, Mpumalanga, Gauteng, North-West and Northern Province	1413.72	microphyll	temperate
Olinia emarginata	Mountain forests, wooden kloofs	South Africa: KwaZulu-Natal, Mpumalanga, Gauteng, North-West and Northern Province	452.39	microphyll	temperate
Olinia vanguerioides	High moutain tips and exposed rocky slopes	South Africa, Swaziland, Moçambique, Uganda, Kenya, Tanzania, Malawi, Rwanda, Ethiopia, Republic of the Congo, Zambia, Angola (scattered populations)	810.53	microphyll	temperate
Penaea acutifolia	Mostly damp sandstone slopes	Cape Floristic Region: SE	23.56	leptophyll	fynbos
Penaea cneorum ssp. gigantea	Damp sandstone slopes and streambanks	Cape Floristic Region: SW, LB, SE	176.71	nanophyll	fynbos
Penaea cneorum ssp. ovata	Damp sandstone slopes and streambanks	Cape Floristic Region: SW, LB, SE	31.42	nanophyll	fynbos
Penaea cneorum ssp. lanceolata	Damp sandstone slopes and streambanks	Cape Floristic Region: SW, LB, SE	24.74	leptophyll	fynbos
Penaea mucronata	Mostly rocky sandstone slopes	Cape Floristic Region: NW, SW, AP, LB	32.99	nanophyll	fynbos
Penaea cneorum ssp. cneorum	Damp sandstone slopes and streambanks	Cape Floristic Region: SW, LB, SE	98.96	nanophyll	fynbos
Penaea cneorum ssp. ruscifolia	Damp sandstone slopes and streambanks	Cape Floristic Region: SW, LB, SE	106.81	nanophyll	fynbos
Brachysiphon microphyllus	Sandstone rocks,	Cape Floristic Region: KM	3.93	leptophyll	fynbos
Penaea dahlgrenii	Sandstone slopes along streams	Cape Floristic Region: LB	18.85	leptophyll	fynbos

Stylapterus ericifolius	Sandstone slopes	Cape Floristic Region: LB	5.89	leptophyll	fynbos
Stylapterus fruticulosus	Sandy flats	Cape Floristic Region: SW	14.14	leptophyll	fynbos
Stylapterus ericoides ssp. pallidus	Mountain streams at low elevations	Cape Floristic Region: SW	10.21	leptophyll	fynbos
Brachysiphon fucatus	Cool, rocky sandstone slopes	Cape Floristic Region: SW	56.55	nanophyll	fynbos
Stylapterus micranthus	Streambanks	Cape Floristic Region: SW	78.54	nanophyll	fynbos
Brachysiphon acutus	Rocky sandstone slopes	Cape Floristic Region: SW	42.41	nanophyll	fynbos
Brachysiphon mundii	Limestone rocks and cliffs	Cape Floristic Region: AP	4.71	leptophyll	fynbos
Saltera sarcocolla	Rocky sandstone slopes	Cape Floristic Region: SW, AP	113.10	nanophyll	fynbos
Sonderothamnus speciosus	Rocky sandstone stlopes	Cape Floristic Region: SW	28.27	nanophyll	fynbos
Sonderothamnus petraeus	Sandstone rocks and cliffs	Cape Floristic Region: SW	28.08	nanophyll	fynbos
Brachysiphon rupestris	Sandstone rocks	Cape Floristic Region: SW	49.48	nanophyll	fynbos
Glischrocolla formosa	Rocks and cliffs at high altitude	Cape Floristic Region: SW	30.63	nanophyll	fynbos
Endonema laterifolia	Rocky sandstone slopes along streams	Cape Floristic Region: SW	188.50	nanophyll	fynbos
Endonema retzioides	Rocky southern sandstone slopes	Cape Floristic Region: SW	47.12	nanophyll	fynbos

Table 6. Divergence time estimates (see Figure 2) and results of the ancestral character state reconstructions (see Figure 5) for selected nodes of interest. Dating results were obtained by analyzing the chloroplast dataset with *multidivtime* (Thorne *et al.*, 1998; Thorne and Kishino, 2002). The age estimates marked with asterisks are from Rutschmann *et al.*, submitted. The Maximum likelihood ancestral character state reconstructions were performed in ANCML (Schluter *et al.*, 1997) for the continuous leaf size character and in Mesquite 1.06 (Maddison and Maddison, 2005) for the discrete leaf size and vegetation type classes, respectively. P values represent the proportional likelihood for the inferred ancestral states. P values for the other states are lower and not shown here, but visible in the pie charts in Figure 5.

Node in Figures 2/5	**Node description**	**Mean age ± standard deviation [mys]**	**95% credibility interval [mys]**	**Reconstucted leaf size (area in mm^2)**	**Reconstructed leaf size class**	**Reconstructed vegetation type class**
a	Crypteroniaceae stem	79.7 ± 7.6*	64.5 – 94.8*	7314	mesophyll, p = 0.38	tropical, p = 0.81
b	Crypteroniaceae crown	50.55 ±8.6*	34.6 – 68.0*	11628	mesophyll, p = 0.70	tropical, p = 0.99
c	Alzateaceae divergence	70.6 ± 7.9*	55.0 – 85.5*	5967	mesophyll, p = 0.32	tropical, p = 0.75
d	Penaeaceae stem	58.4 ± 8.2	45.1 – 76.3	4329	microphyll, p = 0.37	temperate, p = 0.51
e	Penaeaceae crown	16.6 ± 3.6	10.7 – 24.8	605	nanophyll, p = 0.98	fynbos, p = 0.99
f	Oliniaceae stem	53.5 ± 7.9	40.5 – 70.7	3765	microphyll, p = 0.47	temperate, p = 0.79
g	Oliniaceae crown	4.7 ± 2.1	1.5 – 9.5	1485	microphyll, p = 0.98	temperate, p = 0.99

Figures

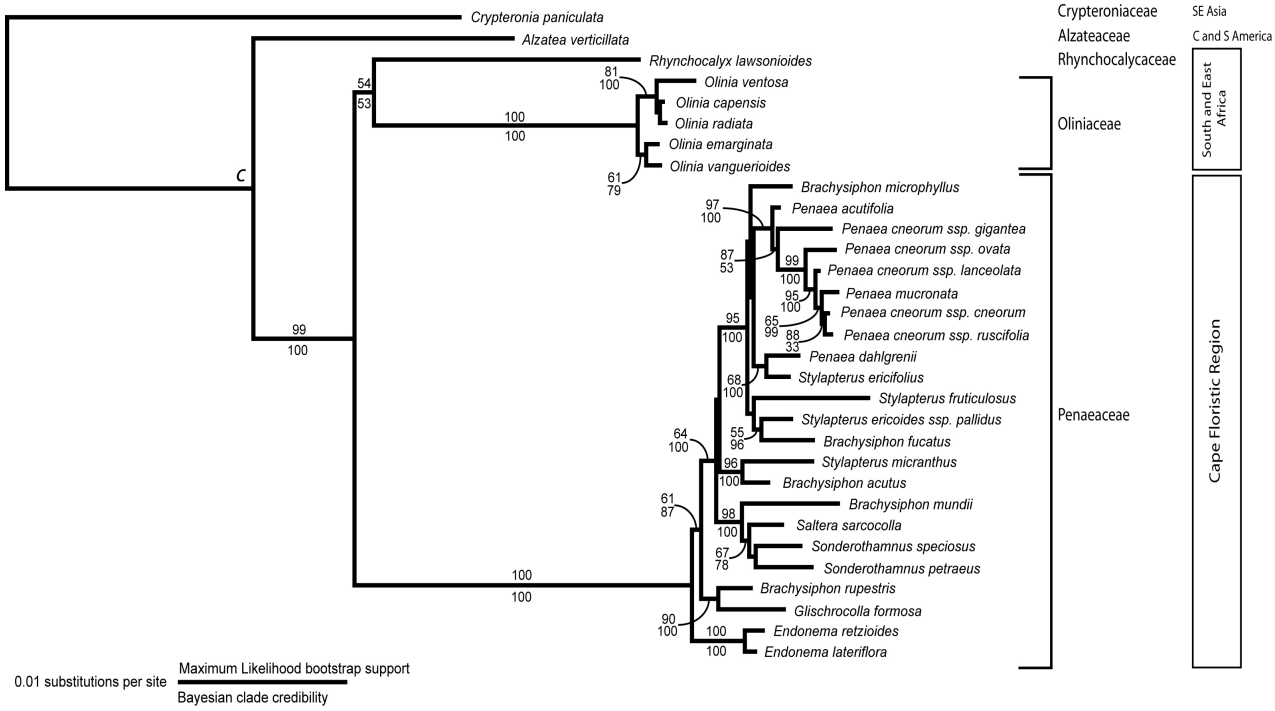


Figure 1a. MrBayes majority rule consensus tree with mean branch lengths, based on eight combined chloroplast gene sequences (*rbc*L, *ndh*F, *rpl*16-intron, *rps*16-intron, *trn*S-*trn*G, *atp*B-*rbc*L spacer, *mat*K, and *psb*A-*trn*H; 6571 characters). Maximum likelihood bootstrap support values and Bayesian clade credibility values are reported above and below the branches, respectively. General distribution ranges are reported to the right of the tree. *Crypteronia paniculata* was used as dating outgroup. Node *c* represents the phylogenetic split between the African Penaeaceae/Oliniaceae and the South American Alzateaceae and was used as a secondary calibration point.

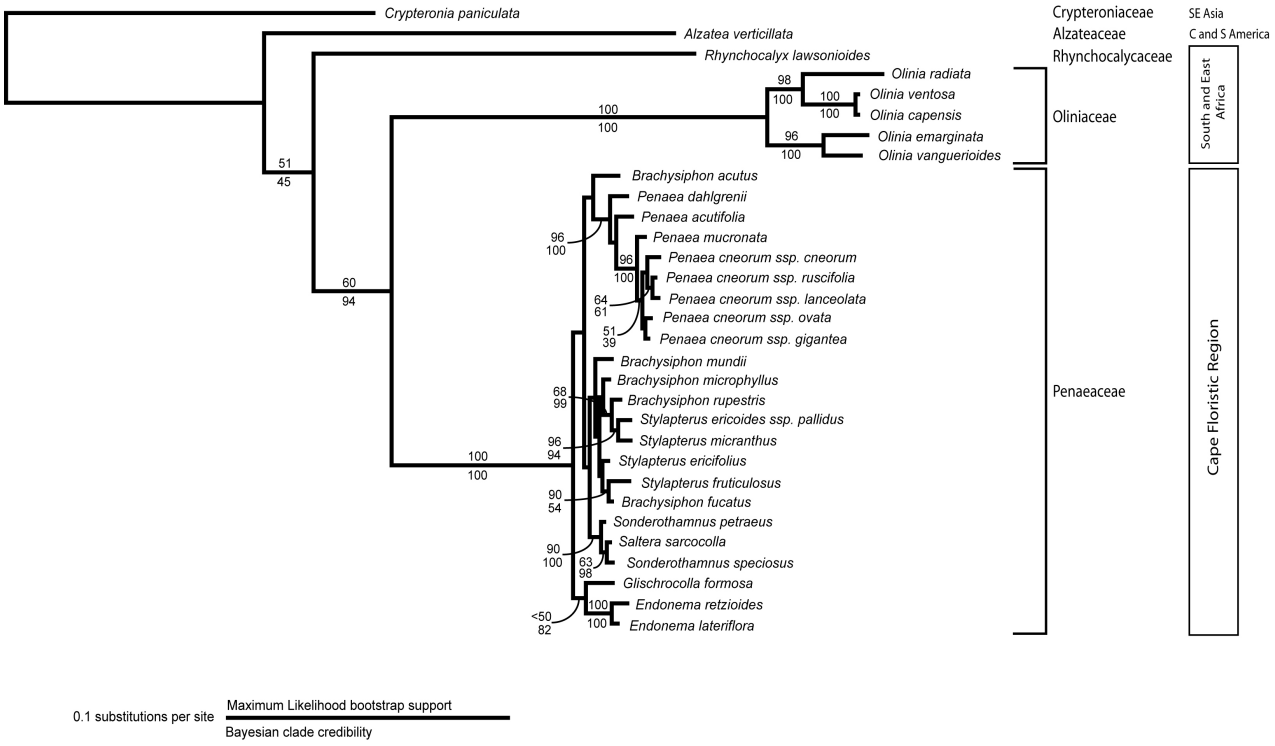


Figure 1b. MrBayes majority rule consensus tree with mean branch lengths, based on three combined nuclear gene sequences (nr18S, nr26S, and ITS; 3384 characters). Maximum likelihood bootstrap support values and Bayesian clade credibility values are reported above and below the branches, respectively. General distribution ranges are reported to the right of the tree.

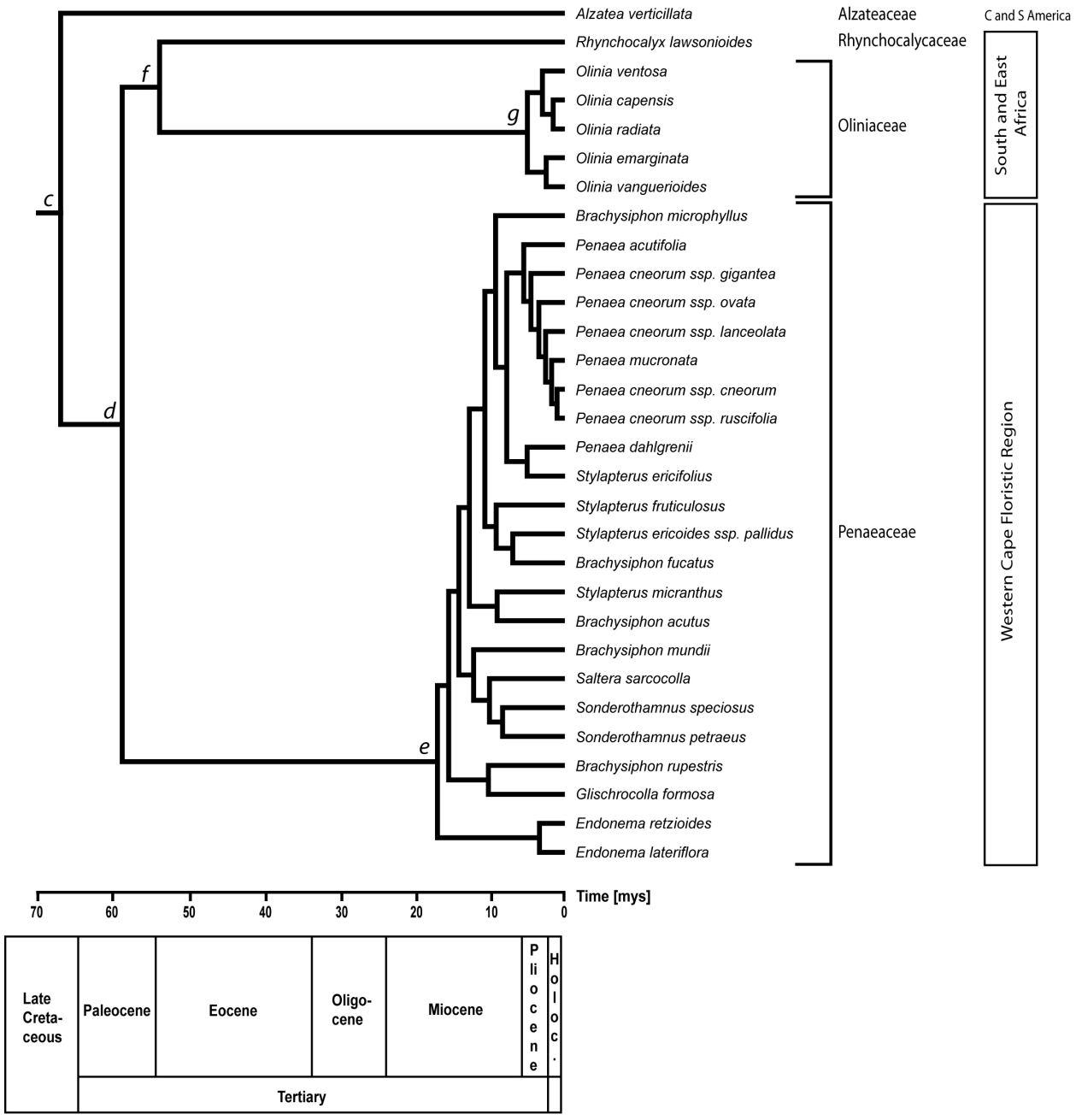


Figure 2. Chronogram obtained by estimating divergence times with *multidivtime* (Thorne *et al.*, 1998) based on the chloroplast dataset. The branch lengths and ages represent posterior mean values. Node *c* represents the phylogenetic split between the African Penaeaceae/Oliniaceae and the South American Alzateaceae. It was used as a secondary calibration point based on earlier age estimates in Rutschmann *et al.*, submitted. The other node labels refer to the nodes of interest used in the ancestral character state reconstruction (see Table 6 and Figure 5). *Crypteronia paniculata* was used as dating outgroup and removed during the analysis (see Figure 1).

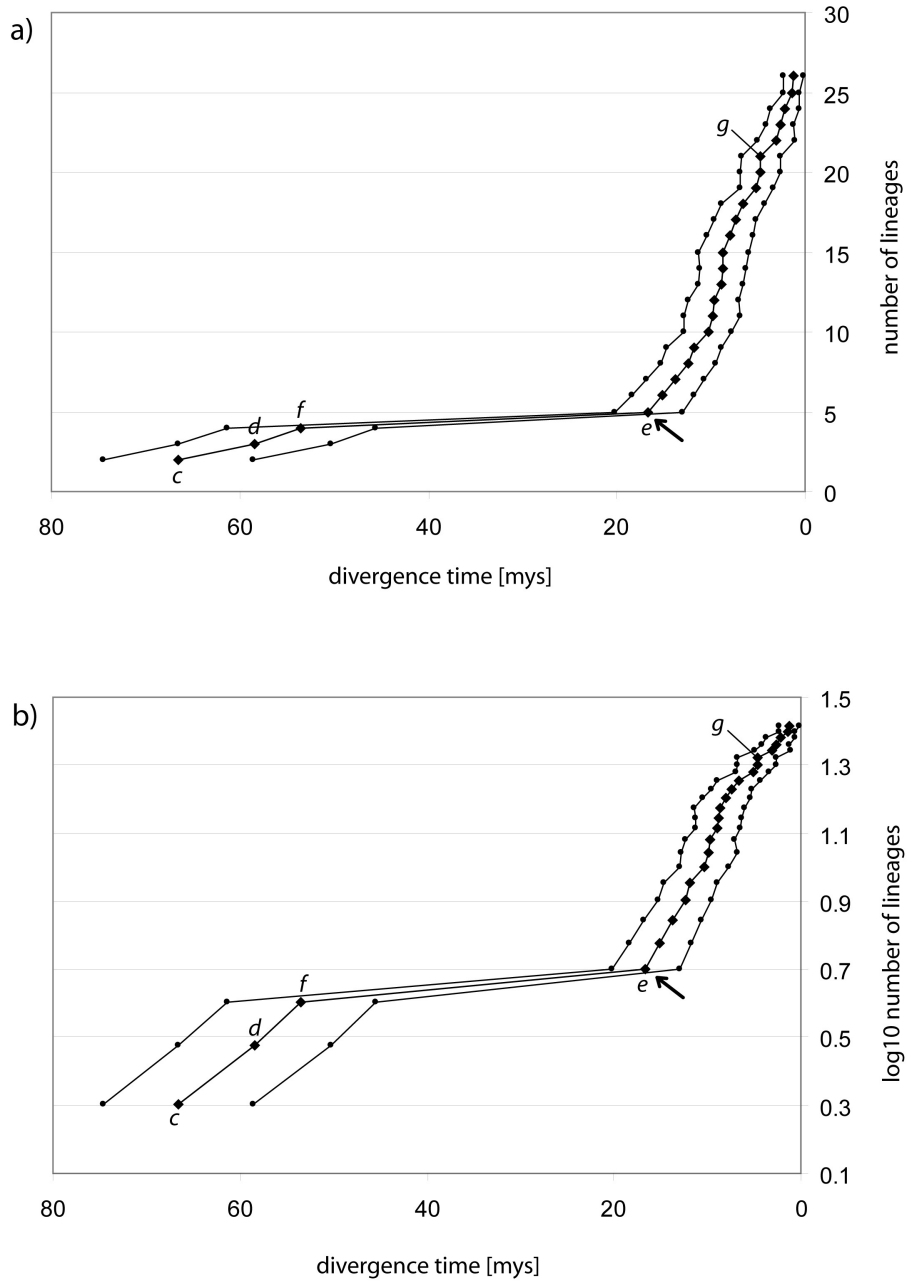


Figure 3. Lineages through time (LTT) plots, obtained by evaluating the chronogram shown in Figure 2 (full dataset), but without subspecies. The number of lineages (Figure 3a) and the log-transformed number of lineages (Figure 3b), respectively, are plotted against estimated divergence times. The data points represent mean estimated ages (diamonds) surrounded by standard deviation values (dots). Data points *c* to *g* represent the nodes of interest used in the ancestral character state reconstructions (see Figures 2 and 5). The arrows mark the point in time (16.6 mys ago) where the diversification rate changes abruptly according to the diversification model selected in the survival analysis (see Table 4; Paradis, 1997).

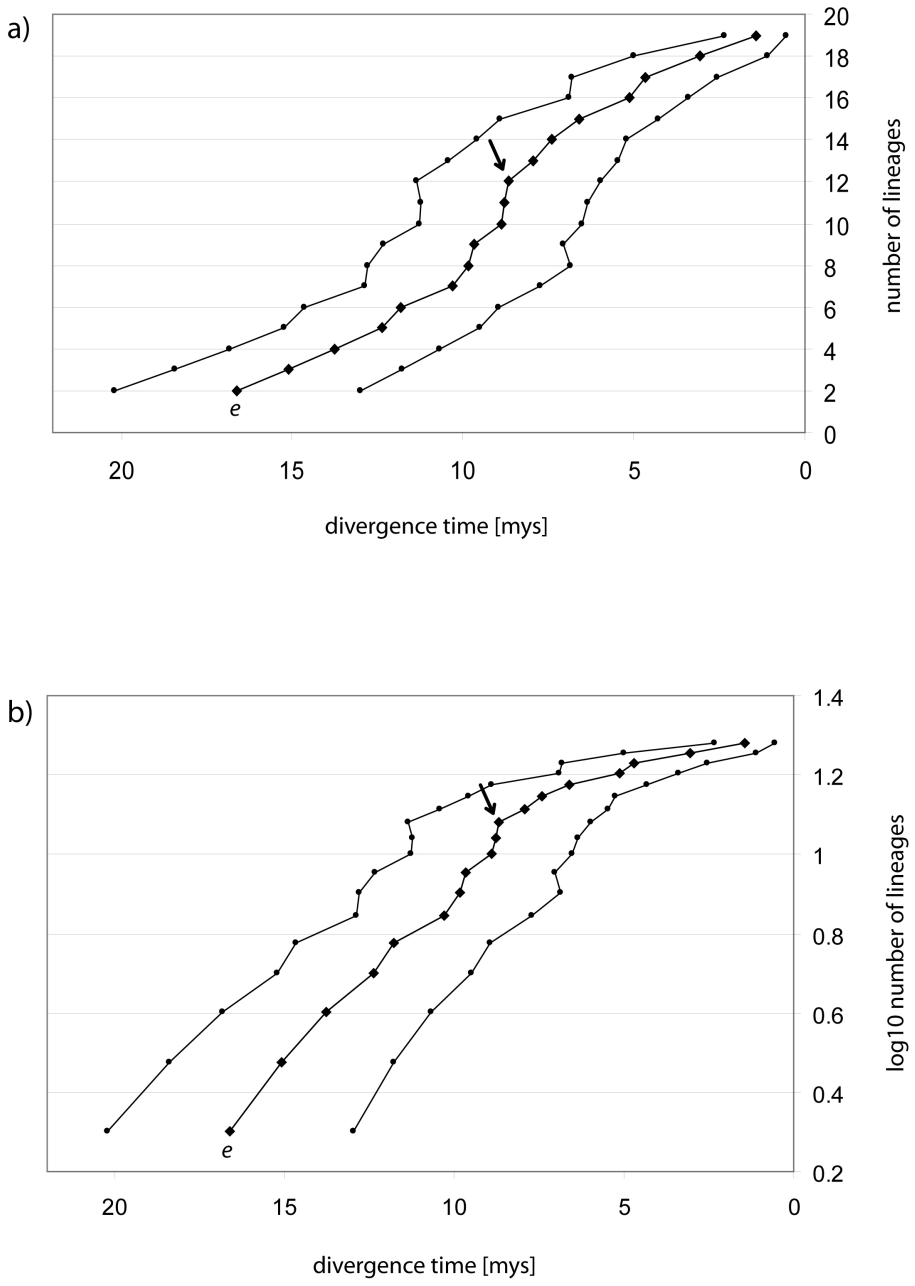


Figure 4. Lineages through time (LTT) plots, obtained by evaluating only the lineages of the Penaeaceae crown group (reduced dataset; see Figure 2), without subspecies. The number of lineages (Figure 4a) and the log-transformed number of lineages (Figure 4b), respectively, are plotted against estimated divergence times. The data points represent mean estimated ages (diamonds) surrounded by standard deviation values (dots). Data point *e* represents the Penaeaceae crown node (see Figures 2 and 5). The arrows mark the point in time (8.7 mys ago) where a supposed change in diversification rate was statistically tested and rejected by the survival analysis (see Table 4; Paradis, 1997).

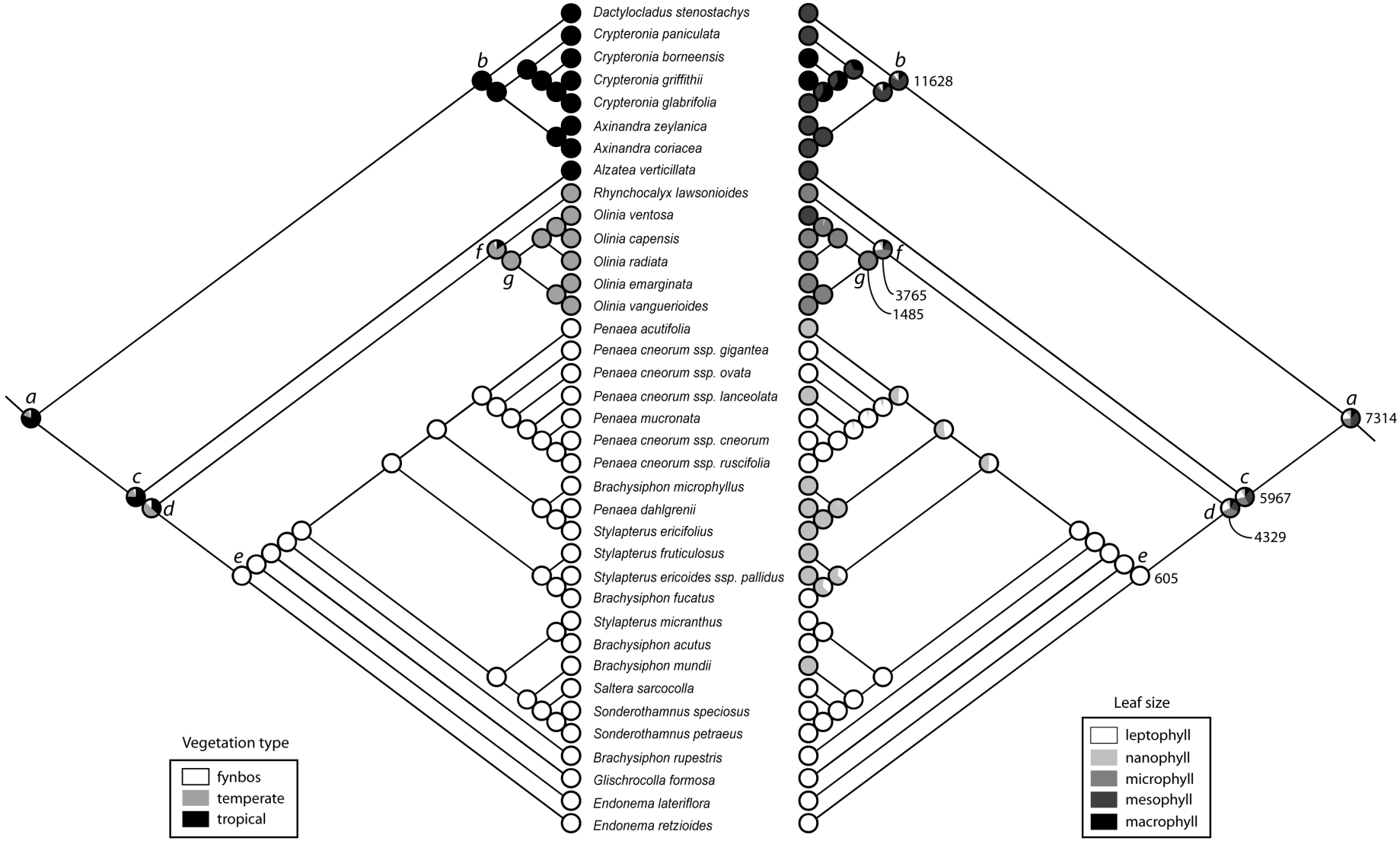


Figure 5 (see legend on next page)

Figure 5 (previous page). Distribution of reconstructed ancestral states for leaf size (left side; three character states) and vegetation type (right side; five character states), optimized under maximum likelihood in Mesquite 1.06 (Maddison and Maddison, 2005) by using the Mk1 model (Table 6). The small pie charts associated to the nodes indicate the proportional likelihoods for the different discrete character states. For the nodes of interest, labelled *a* to *g*, more detailed results are reported in Table 6. The numbers associated with some nodes in the tree to the right represent the continuous leaf size values reconstructed by using ANCML (Schluter *et al.*, 1997; see Table 6).

Chapter V.

Assessing calibration uncertainty in molecular dating: The assignment of fossils to alternative calibration points

Frank Rutschmann, Torsten Eriksson[1], Kamariah Abu Salim[2], and Elena Conti[3]

[1]Bergius Foundation, Royal Swedish Academy of Sciences, SE-104 05 Stockholm, Sweden

[2]Universiti Brunei Darussalam, Department of Biology, Jalan Tungku Link, Gadong BE 1410, Negara Brunei Darussalam

[3]Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland

Submitted for publication in

Systematic Biology

Abstract

While recent methodological advances have allowed the incorporation of rate variation in molecular dating analyses, the calibration procedure, performed mainly through fossils, has proved more resistant to amelioration. One source of uncertainty pertains to the assignment of fossils to specific nodes in a phylogeny, especially when alternative possibilities exist that can be equally justified on morphological grounds. Here we expand on a recently developed fossil cross-validation method to evaluate whether alternative nodal assignments produce calibration sets that differ in their internal consistency. We use an enlarged Crypteroniaceae-centered phylogeny of Myrtales, six fossils, and 72 combinations of calibration points, termed calibration sets, to identify (i) the fossil assignments that produce the most internally consistent calibration sets and (ii) the overall mean age, derived from all calibration sets, for the split of the Southeast Asian Crypteroniaceae from their West Gondwanan sister clade (node X). We found that a correlation exists between *s* values, devised to measure the consistency among the calibration points of a calibration set, and nodal distances among calibration points. By ranking all sets according to the percent deviation of *s* from the regression line with nodal distance, we identified the sets with the highest level of corrected calibration-set consistency. These sets generated lower standard deviations associated with the age of node X than sets characterized by lower corrected consistency. The three calibration sets with the highest corrected consistencies produced age estimates for node X of 79.70 mys, 79.14 mys, and 78.15 mys. The overall mean age, derived from all 72 calibration sets, was 77.74 mys, with a 95% credibility interval of 60.89 to 94.21 mys. These timeframes are most compatible with the hypothesis that the Crypteroniaceae stem lineage dispersed from Africa to the Deccan plate as it drifted northward during the Late Cretaceous.

Key words: Molecular dating, divergence times, fossil calibration, calibration point, cross-validation, Crypteroniaceae, biogeography, out-of-India.

Introduction

The use of DNA sequences to estimate the timing of evolutionary events is increasingly popular. Based on the central idea that the differences between the DNA sequences of two species is a function of the time since their evolutionary separation (Zuckerkandl and Pauling, 1965), molecular dating has been used as a method to investigate both patterns and processes of evolution (Sanderson *et al.*, 2004; Magallón, 2004; Welch and Bromham, 2005; Renner, 2005; Rutschmann, 2006).

However, significant methodological challenges affect the use of molecular dating approaches. While recent studies have addressed the issue of variation among substitution rates (Sanderson, 1997, 2002; Thorne *et al.*, 1998; Kishino *et al.*, 2001; Aris-Brosou and Yang, 2002; Yang, 2004; Penny, 2005; Ho and Larson, 2006; Rutschmann, 2006), other difficulties persist, especially concerning the calibration procedure (Lee, 1999; Conti *et al.*, 2004; Magallón, 2004; Reisz and Müller, 2004). Calibration consists in the incorporation of independent (non-molecular) chronological information in a phylogeny to transform relative into absolute divergence times. This information can be based on geological events (e.g., patterns of continental drift, origin of islands and mountain chains) and/or the paleontological record (fossils). Geological calibrations points are assigned to phylogenetic nodes based on the assumption that a geographic barrier caused phylogenetic divergence, thus generating the risk of circular reasoning, if the chronogram derived from the calibration is used to test biogeographical scenarios (Conti *et al.*, 2004; Magallón, 2004). Nevertheless, geological events can provide important validation of dating estimates produced with other types of calibration (e.g., Conti *et al.*, 2002; Sytsma *et al.*, 2004; Bell and Donoghue, 2005).

While the fossil record is widely regarded as the best source of non-molecular information about the ages of selected clades (Marshall, 1990b; Sanderson, 1998; Magallón and Sanderson, 2001), several interrelated problems plague its use for calibration purposes, including (i) erroneous fossil age estimates, (ii) the incompleteness of the fossil record, (iii) the assignment of fossils to specific nodes in a phylogeny, and (iv) the number of fossils used for calibration. The first problem may depend on misleading stratigraphic correlations or improper radiometric dating (Conroy and van Tuinen, 2003) and can only be addressed by improving the geological dating procedures.

The incompleteness of the fossil record may result from the failure of entire species to be preserved or discovered as fossils (Darwin, 1859). This lack of information makes it difficult or impossible to estimate the temporal gaps between the divergence of two lineages, the origin of a synapomorphy, and the discovery of that synapomorphy in the fossil record (see Figure 1 in Magallón, 2004; Springer, 1995; Foote and Sepkoski, 1999). Fossils can thus provide only minimum ages for any lineage. Recently developed methods aimed at estimating the gap between

the time of first appearance of a synapomorphy in the fossil record and the time of divergence between two lineages are based on the idea that the size of the gap is inversely correlated with the quality and density of the fossil record within a given stratigraphic interval, and dependent on the rates of origination, extinction and preservation of the focus lineage (Marshall, 1990a, 1990b; Foote *et al.*, 1999a, 1999b, Tavaré *et al.*, 2002; Yang and Rannala, 2005).

Depending on their preservation state, relative abundance, and the distinctiveness of selected morphological traits, it can be problematic to unambiguously assign fossils to a particular clade in a given phylogeny (Benton and Ayala, 2003; Doyle and Donoghue, 1993). More specifically, it is necessary to determine whether the fossil represents an extinct member of the stem or the crown group of extant taxa (Hennig, 1969; de Queiroz and Gauthier, 1990; Doyle and Donoghue, 1993; Magallón and Sanderson, 2001; Magallón, 2004). Ideally, the assignment would be based on a comprehensive cladistic morphological analysis of both extant and extinct taxa, but, due to their complexity, such analyses remain regrettably rare (Conti *et al.*, 2004; Near *et al.*, 2005b). In practice, assignment of fossils to selected nodes (called "calibration nodes" from now on) is usually based on more intuitive comparisons between the character states of the fossil and the distribution of synapomorphies in the phylogeny. When a fossil is finally attached to a node, the node in question assumes the age of the fossil, thus becoming a "calibration point".

For the reasons explained above, the use of a single fossil for calibration can produce strongly biased molecular age estimates (Alroy, 1999; Lee, 1999; Smith and Peterson, 2002; Conroy and van Tuinen, 2003; Hedges and Kumar, 2003; van Tuinen and Hadly, 2003; Graur and Martin, 2004; Reisz and Müller, 2004). Additionally, the nodal distance of the calibration point to the node(s) of interest and the root of the phylogeny may strongly influence the estimated ages (Smith and Peterson, 2002; Conroy and van Tuinen, 2003; Reisz and Müller, 2004). Therefore, it seems desirable to use multiple fossils, preferably placed in different clades (Brochu, 2004), for molecular dating purposes, in the hope that the biases built into their assignment to specific calibration nodes may cancel each other out (Smith and Peterson, 2002; Soltis *et al.*, 2002; Conroy and van Tuinen, 2003).

Recently developed methods allow for the incorporation of multiple calibration points (termed "multi-calibration" from now on) in the dating procedure (Sanderson, 1997, 2002; Thorne *et al.*, 1998; Kishino *et al.*, 2001; Yang, 2004; Yang and Rannala, 2005; Drummond *et al.*, 2006). When multiple fossils are available, it is possible to use one fossil at a time to generate age estimates for the nodes to which the other fossils are assigned and then compare the estimated ages with the fossil ages at those nodes, essentially leading to an assessment of the consistency among calibration points (Near and Sanderson, 2004; Near *et al.*, 2005b). This procedure, known as fossil cross-validation, allows for the identification and removal of incongruent calibration point(s), and

has been applied in molecular dating studies of monocotyledons (where two out of eight fossils were removed; Near and Sanderson, 2004), placental mammals (two out of nine; Near and Sanderson, 2004), turtles (seven out of 17; Near *et al.*, 2005b), centrarchid fishes (four out of 10; Near *et al.*, 2005a) and decapods (no fossils removed; Porter *et al.*, 2005). To summarize, Near and Sanderson (2004) devised a method to assess whether individual calibration points were inconsistent with the other points of a calibration set, leading to the removal of the corresponding fossil(s) from the calibration procedure. In the study presented here, we expand their approach to evaluate whether alternative assignments of available fossils to different calibration nodes produce more or less internally consistent calibration sets.

To examine the calibration problem mentioned above, we utilize a Myrtales data set centered on the relationships of Crypteroniaceae and related families (named "Crypteroniaceae phylogeny" from now on). Earlier molecular dating results (Conti *et al.*, 2002, 2004; Rutschmann *et al.*, 2004) suggested a possible Gondwanan origin of these families in the Early to Middle Cretaceous, followed by the dispersal of the Crypteroniaceae stem lineage to the Deccan plate (comprising India and Madagascar) while it was rafting along the African coast, ca. 125 to 84 mys ago (McLoughlin, 2001; Plummer and Belle, 1995; Storey *et al.*, 1995), a biogeographic scenario known as the out-of-India hypothesis (McKenna, 1973; Ashton and Gunatilleke, 1987; Macey *et al.*, 2000; Morley, 2000; Bossuyt and Milinkovitch, 2001). However, the estimated age of the crucial biogeographic node, representing the split between the Southeast Asian Crypteroniaceae and the African/South American sister clade (node X; Figure 1), ranged from 106 to 141 mys (Conti *et al.*, 2002), 62 to 109 mys (Rutschmann *et al.*, 2004), 57 to 79 mys (Moyle, 2004), and 52 mys (Sytsma et al, 2004), depending on gene and taxon sampling, but mostly on the contrasting assignment of selected fossils to different nodes in the Myrtales phylogeny (Conti *et al.*, 2004; Moyle, 2004). Given the controversial nature of calibration, the contradicting calibration procedures previously applied to date the Crypteroniaceae phylogeny, and the availability of multiple fossils in Myrtales, this group of taxa provides an ideal case study to investigate problems of calibration, at the same time attempting to refine the age estimates that are central to the biogeographic history of Crypteroniaceae.

To calibrate the Crypteroniaceae phylogeny we use six fossils. Based on the morphological traits preserved in the fossils, five out of the six fossils can each be assigned to two or three different nodes in the phylogeny (Figure 1; see below). In total, 72 different combinations of six calibration points are possible, each combination forming a calibration set. Here, we employ an expanded molecular data set and the six fossils to address the following questions: 1) How can the fossil cross-validation procedure be used to assign a fossil to a calibration point that is most internally consistent with the other points in a calibration set? In other words, which fossil

assignments produce the calibration sets that are most internally consistent? 2) What is the distribution of ages and the overall mean age for the split between the Southeast Asian Crypteroniaceae and their West Gondwanan sister clade that is estimated from using all 72 calibration sets?

To our knowledge, this is the first study that explores the application of fossil cross-validation to the problem of nodal assignment for selected fossils (termed "fossil nodal assignment" from now on). While we realize that the results presented here by no means represent a panacea to the complex challenges of calibration (Wray, 2001; Magallón, 2004; Near and Sanderson, 2004; Reisz and Müller, 2004; Müller and Reisz, 2005), we nevertheless try to present a practical approach, based on available methods, to address difficulties encountered in most molecular dating analyses, but all too often ignored.

Materials and Methods

Taxon and DNA sampling

Sampling was expanded from published studies (Conti *et al.*, 2002; Schönenberger and Conti, 2003; Rutschmann *et al.*, 2004) to include DNA sequences from three plastid (*rbc*L, *ndh*F and *rpl*16-intron) and two nuclear (ribosomal 18S and 26S; termed nr18S and nr26S from now on) loci for 74 taxa (see Table 1). In total, 270 new sequences were generated for this study. Almost complete taxon sampling was achieved for Crypteroniaceae (seven of 12 species; Mentink and Baas, 1992; Pereira and Wong, 1995; Pereira, 1996), Alzateaceae (one of one species; Graham, 1984), Rhynchocalycaceae (one of one species; Johnson and Briggs, 1984), Penaeaceae (19 of 23 species, plus four subspecies of *Penaea cneorum*; Dahlgren and Thorne, 1984; Dahlgren and van Wyk, 1988), and Oliniaceae (five of eight species; Tobe and Raven, 1984; Sebola and Balkwill, 1999). The six missing taxa of Crypteroniaceae, *Axinandra alata, A. beccariana*, *Crypteronia elegans*, *C.macrophylla*, *C. cummingii*, and *C. paniculata var. affinis* are either very rare, known only from the type, or collected in now deforested areas, for ex., in Kalimantan (Indonesia; Pereira, 1996). DNA extractions from dried vouchers were unsuccessful, because all herbarium specimens of these taxa were treated with ethanol after collection. The four missing Penaeaceae, *Stylapterus barbatus*, *S. dubius*, *S. sulcatus*, and *S. candolleanus* are also either very rare, occur only in restricted areas, or were collected only once or twice each (pers. comm. Jürg Schönenberger). Following Sebola and Balkwill (1999), the three missing Oliniaceae are *Olinia discolor*, *O. rochetiana,* and *O. micrantha*. The sampling of the remaining taxa was designed to assign the fossils as precisely as possible and to represent clade heterogeneity at the family level (Melastomataceae, Myrtaceae s.l., Vochysiaceae, Onagraceae, Lythraceae), based on published phylogenies (Conti *et al.*, 1996, 1997; Renner, 2004; Sytsma *et al.*, 2004). Phylogenies were rooted using representatives of Lythraceae, i.e., *Duabanga grandiflora* and *Cuphea hyssopifolia*, based on the results of global Myrtales analyses (Conti *et al.*, 1996, 1997; Sytsma *et al.*, 2004).

DNA Extractions, PCR, sequencing, and alignment

DNA was extracted as described in Schönenberger and Conti (2003) and Rutschmann *et al.* (2004). Primers from Zurawski *et al.* (1981), Olmstead and Sweere (1994), Baum *et al.* (1998), Bult *et al.* (1992), and Kuzoff *et al.* (1998) were used to amplify and sequence *rbc*L, *ndh*F, *rpl*16-intron, nr18S, and nr26S, respectively. PCR and sequencing procedures followed the protocols described in Rutschmann *et al.* (2004). The software Sequencher 4.5 (Gene Codes, Ann Arbor, MI, USA) was used to edit, assemble, and proof-read contigs for complementary strands. *Rbc*L, nr26S and nr18S

sequences were readily aligned by eye, while *ndh*F and *rpl*16-intron sequences were first aligned using Clustal X 1.83 (Thompson *et al.*, 1997) prior to final visual adjustment with MacClade 4.07 (Maddison and Maddison, 2000).

Phylogenetic analyses

Tree topology and branch lengths were estimated using MrBayes version 3.0b4 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Plastid (*rbc*L, *ndh*F, and *rpl*16-intron) and nuclear (nr18S, nr26S) partitions were first analyzed separately (results not shown). Because the respective 80% majority rule consensus bootstrap trees had no well supported incongruencies, both datasets were combined.

Model selection for the combined data set was performed in MrAIC 1.4 (Nylander, 2005b), a program that uses PHYML 2.4.4 (Guindon and Gascuel, 2003) to find the maximum of the likelihood function under 24 models of molecular evolution. MrAIC identified the optimal models according to two different selection criteria: the corrected Akaike information criterion (AICc) and the Bayesian information criterion (BIC). Parameter values for the chosen models were then estimated simultaneously for each partition during topology and branch length optimization in MrBayes.

Bayesian topology estimation used one cold and three incrementally heated Markov chain Monte Carlo chains (MCMC) run for 1.5 x 10^6 cycles, with trees sampled every 100^{th} generation, each using a random tree as a starting point and the default temperature parameter value of 0.2. For each data set, MCMC runs were repeated twice. The first 5000 trees were discarded as burn-in after checking for stationarity on the logarithmic likelihoods curves. The remaining trees were used to construct one Bayesian consensus tree with mean branch lengths and to calculate the clade credibility values (Figure 1).

Statistical support for individual branches was also calculated by bootstrap resampling using the Perl script BootPHYML 3.4 (Nylander, 2005a). This program first generates 1000 pseudo-replicates in SEQBOOT (part of Phylip 3.63; Felsenstein, 2004), then performs a maximum likelihood analysis in PHYML (Guindon and Gascuel, 2003) for each replicate under the selected model of evolution, and finally computes a 50% majority rule consensus tree by using CONSENSE (also part of the Phylip package).

Evaluation of rate heterogeneity

The constancy of nucleotide substitution rates was checked using the χ^2 molecular clock test of Langley and Fitch (1974; LF χ^2 test, implemented in *r8s* 1.7; Sanderson, 2003), which compares the observed branch lengths with the branch lengths predicted under a clock model.

Fossil nodal assignment

After carefully reviewing the paleobotanic literature of Myrtales, six fossils were selected for calibration in molecular dating analyses (see Table 2 and Figure 1). Based both on their morphological characters and previous nodal assignments in published phylogenies (Melastomataceae: Clausing and Renner, 2001; Renner *et al.*, 2001; Renner and Meyer, 2001; Renner, 2004; Myrtaceae: Sytsma *et al.*, 2004; Onagraceae: Berry *et al.*, 2004), five out of the six fossils could justifiably be assigned to different nodes.

Fossil 1. Fossil 1 is represented by leaves from the Early Eocene (53 mys) of North Dakota (Table 2). Hickey (1977), who first described them, stated that they resemble most closely the leaves of extant Miconieae and Merianieae, but assessed that "they all differ, however, in not being deeply cordate and in having tertiaries which do not form a good V pattern". Renner *et al.* (2001) further confirmed the resemblance with leaves of modern Miconieae and Merianieae. However, because all Melastomataceae, including *Pternandra,* share the same basic kind of acrodromous leaf venation, Renner *et al.* (2001) decided to assign the fossils to the entire Melastomataceae crown group (node *1a* in the present study; see Figure 1), a view later shared by Sytsma *et al.* (2004) in a Myrtales-wide dating analysis. Conversely, in the light of the stated similarities with Miconieae and Merianieae and the known temporal gap between lineage divergence and first appearance of a synapmorphy in the fossil record, Conti *et al.* (2002), Morley and Dick (2003) and Rutschmann *et al.* (2004) decided to assign the fossil leaves either to the stem (node *1b*) or to the crown group (node *1c*) that includes Miconieae and Merianieae (Figure 1). Renner (2004) included both assignment possibilities in subsequent Bayesian dating analyses of Melastomataceae. In summary, the Miocene leaves of Melastomataceae can be defensibly assigned to three different nodes (*1a*, *1b*, and *1c*; see Figure 1). Because the branch separating node *1a* and *1b* is long, fossil assignments above or below this branch was expected to strongly influence the dating results.

Fossil 2. Fossil 2 is represented by fossil seeds from Miocene deposits in Siberia, the Tambov region, Belarus, Poland, several sites in Germany, and Belgium (23-26 mys, Dorofeev, 1960, 1963, 1988; Collinson and Pingen, 1992; Dyjor *et al.*, 1992; Fairon-Demaret, 1994; Mai, 1995, 2000; Table 2). These seeds are most similar to those of extant members of Osbeckieae and Rhexieae (now Melastomeae; Clausing and Renner, 2001), but differ in several significant features, especially the presence of multicellular tubercles (Collinson and Pingen, 1992). Renner and Meyer (2001) assigned the fossils to the crown group of Melastomeae, because "this kind of testa ornamentation is synapomorphic for the *Rhexia-Arthrostemma-Pachyloma* subclade of Melastomeae". Renner *et al.* (2001), Conti *et al.* (2002), Rutschmann *et al.* (2004), Renner (2004) and Sytsma *et al.* (2004) all followed this interpretation, although it is difficult to establish with

certainty whether these fossils should be assigned to the base of the Melastomeae crown group or to more recent nodes in the tribe. Given our current taxon sampling, the only possible assignment of fossil 2 was to the crown group of Melastomeae, as in previous studies (Figure 1).

Fossil 3. Fossil 3 is represented by the pollen *Myrtaceidites lisamae* (syn. *Syncolporites lisamae*) from the Santonian of Gabon (86 mys; Herngreen, 1975; Boltenhagen, 1976; Muller, 1981), the lower Senonian of Borneo (89 – 83.5 mys; Muller, 1968), and the Maastrichtian of Colombia (71.3 – 65 mys; van der Hammen, 1954; Table 2). Sytsma *et al.* (2004) attributed this pollen to the crown group of Myrtaceae s.s. (corresponding to node *3c* in Figure 1), but pollen grains of Myrtaceae s.s. and Psiloxyloideae (Wilson *et al.*, 2005) are difficult to distinguish. Therefore, a minimal age of 86 mys can be assigned to three different nodes: *3a* (stem lineage of Myrtaceae s.l.), *3b* (crown group of Myrtaceae s.l. or stem lineage of Myrtaceae s.s.), and node *3c* (crown group of Myrtaceae s.s.; Figure 1). Again, because the branch separating nodes *3a* and *3b* is long, different assignments were expected to have a relevant impact on the dating results.

Fossil 4. Fossil 4 is represented by fruits and seeds of *Paleomyrtinaea* from the late Paleocene of North Dakota (56 mys; Crane *et al.*, 1990; Pigg *et al.*, 1993) and the early Eocene of British Columbia (54 mys; Manchester, 1999; Table 2). These fruits and seeds have seed coat features (Lantern and Sharp, 1989) and an unornamented C-shape embryo (Lantern and Stevenson, 1986) that resemble those of the largely American subtribe Myrtinae, included in Myrteae (Pigg *et al.*, 1993). Sytsma *et al.* (2004) assigned the fossil to the crown group of the Myrteae (corresponding to node *4b* in Figure 1), because Myrtinae proved to be paraphyletic in their phylogenetic analysis. As our taxon sampling includes two members of Myrteae that do not represent subtribe Myrtinae, we decided to assign fossil 4 either to the Myrteae crown group (node *4b*; as in Sytsma *et al.*, 2004) or the Myrteae stem lineage (node *4a*; Figure 1). The branch separating nodes *4a* and *4b* is short, thus alternative calibrations were not expected to have large effects on nodal age estimates.

Fossil 5. Fossil 5 is represented by the earliest Eucalypt-like fruits from the Middle Eocene Redbank Plains Formation in South Eastern Queensland, Australia (Rozefelds, 1996; discovered by Robert Knezour in 1990; pers. comm. David Greenwood, Brandon University, Manitoba, Canada; Table 2). Similar fossil fruits were found in the Middle Eocene sediments near lake Eyre, Nelly Creek, in northern South Australia (Christophel *et al.*, 1992). They provide a minimum age of 48 mys for the most recent common ancestor (MRCA) of *Eucalyptus* and *Angophora* (node *5b*). Alternatively, we assigned the fossil to the stem lineage of Eucalypteae (node *5a*; Figure 1), because the observed type of capsule is also similar to the fruit of other taxa in Eucalypteae (Rozefelds, 1996). As for fossil 4, the branch separating nodes *5a* and *5b* is short, thus alternative calibrations were not expected to have large effects on nodal age estimates.

Fossil 6. Fossil 6 is represented by the pollen *Diporites aspis* from the Early Oligocene (33.7-28.5 mys) of Otway Basin (Australia), described by Berry *et al.* (1990; Table 2). Comparisons of the fossil pollen with extant members of *Fuchsia* (Daghlian *et al.*, 1985) left no question that *Diporites aspis* represents pollen of *Fuchsia*. Molecular clock analyses of a *Fuchsia* phylogeny calibrated with non-*Fuchsia* fossils resulted in an estimated age interval for the *Fuchsia* crown group that was consistent with the paleobotanical age of *Diporites aspis* (Berry *et al.*, 2004; Sytsma *et al.*, 2004). We therefore assigned a minimal age of 28.5 mys to the *Fuchsia* crown group (node *6b*). Alternatively, we assigned the age of the fossil to the *Fuchsia* stem lineage (node *6a*; Figure 1), because the synapomorphies visible in the fossil pollen might have evolved before the diversification of *Fuchsia*, and closer to the stem node representing the phylogenetic split between *Fuchsia* and its sister clade. As for fossils 4 and 5, the branch separating nodes *6a* and *6b* is short, thus alternative calibrations were not expected to have large effects on nodal age estimates.

Considering all possible assignments of the six fossils reviewed above, 72 different combinations are possible, corresponding to 72 calibration sets, each comprising six calibration points (Table 3). The ages of the six fossils used in each set were assigned as minimal constraints to the corresponding nodes in all molecular dating analyses.

Molecular dating analyses

All dating analyses described below were performed with a Bayesian approach (*multidivtime*; Thorne *et al.*, 1998; Thorne and Kishino, 2002), which uses an MCMC procedure to derive the posterior distributions of rates and times and allows for multiple calibration windows. The analytical procedure followed the steps described in Rutschmann (2005) and was performed on a 3 GHz Pentium IV machine running Ubuntu Linux 5.10. The values of the prior distributions in the last *multidivtime* step (Kishino *et al.*, 2001; Thorne and Kishino, 2002) were specified in units of 10 mys, as suggested in the manual: RTTM = 12, RTTMSD = 3, RTRATE and RTRATESD = 0.00815, BROWNMEAN and BROWNSD = 0.0833, and BIGTIME = 13. The first two values, which flexibly constrain the age of the root, were chosen in light of published estimates for the age of the Myrtales crown group, including 100 – 107 mys (Wikström *et al.*, 2001, 2003), about 105 mys (Magallón and Sanderson, 2005), and 111 mys (Sytsma *et al.*, 2004; see also Sanderson *et al.*, 2004).

We ran the Markov chain for at least 5×10^5 cycles and collected one sample every 100 cycles, without sampling the first 8×10^4 cycles (burn-in sector). Initial experiments with 2×10^6 cycles showed no differences, leading us to the conclusion that convergence was reached much earlier. We performed each analysis at least twice with different initial conditions and checked the output sample files to assure convergence of the Markov chain by using the program Tracer 1.3 (Rambaut and Drummond, 2005), even though we realize that it is not possible to establish with

certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution (Cowles and Carlin, 1996).

Finding the most internally consistent calibration sets by using fossil cross-validation

To evaluate whether some of the 72 calibration sets were more internally consistent than others, we implemented the fossil cross-validation procedure of Near and Sanderson (2004) and Near *et al.* (2005b). While these authors developed the method to identify possibly inconsistent points in a single calibration set, we expanded their approach to compare the internal consistencies of different calibration sets generated by alternative nodal assignments of multiple fossils. Therefore, for each of the 72 calibration sets in our case study, we performed the following steps:

1. We fixed one out of the six calibration points and estimated the ages of the remaining five, unconstrained nodes. This procedure is called "fossil calibration run" from now on.

2. We then calculated the difference D_i between the estimated and the fossil ages of the five unconstrained nodes. This difference was defined by Near and Sanderson (2004) as an absolute deviation measure D_i = (estimated age – fossil age). Instead, we used the relative deviation measure D_i = (estimated age – fossil age)/fossil age), as suggested in Near *et al.* (2005b).

3. Then, we calculated $SS\chi$, the sum of the six squared D_i values (by using equation 2.2 in Near and Sanderson, 2004).

4. The same procedure (steps 1-3) was then repeated for the remaining five fossil calibration runs, obtained by fixing a different calibration point.

5. Based on the six $SS\chi$ scores obtained for all six calibration runs, we calculated the average squared deviation *s* for the calibration set (with equation 2.3 in Near and Sanderson, 2004; see Table 3).

By using this procedure iteratively, we obtained *s* values for all 72 calibration sets. High values of *s* would indicate that one or more calibration points in a set are inconsistent with the others, suggesting that the corresponding fossils were erroneously assigned to the respective nodes, whereas low *s* values would characterize calibration sets with high internal consistency.

In order to account for possible effects related to the molecular dating method (in our case *multidivtime*), the cross-validation experiments were repeated by using penalized likelihood (Sanderson, 2002) implemented in *r8s* (Sanderson, 2003; data not shown). For batch processing the calculations of $SS\chi$ and *s*, we wrote a collection of Perl scripts (available from the first author upon request).

Relationship between average squared deviation s and nodal distance. Because the fossil cross-validation procedure might be influenced by the position of the calibration points

relative to each other (Near *et al.*, 2005a), we tested whether the average squared deviations *s* of the 72 calibration sets were correlated with the number of nodes separating the points of a calibration set (nodal distance; Figure 2). For each calibration set, nodal distances were calculated by first counting the number of nodes between the fixed node and each of the five unconstrained nodes in each fossil calibration run (see above), then summing them up over all runs (see Table 3). Correlation significance was tested by using the F-test statistic under a linear regression model in R (R Development Core Team, 2004) and the Spearman's Rank Correlation test.

Because the 72 calibration sets differed in their degree of correlation between average squared deviation *s* and nodal distance (Figure 2), we also calculated the percent deviation of *s* from the regression line with nodal distance for each calibration set (see Table 3) and plotted the results as a histogram (Figure 3). The percent deviation of *s* represents a corrected measure of internal consistency among the calibration points of a set, termed "corrected calibration-set consistency" from now on.

Effect of corrected calibration-set consistency on dating precision. In order to check whether there is a relationship between the corrected calibration-set consistency and the precision of the dating estimates for the node of interest (node X; see Figure 1), we plotted the percent standard deviation for the age of node X calculated for each of the 72 calibration sets (see below) against the percent deviation of *s* from the regression line with nodal distance (Table 3; Figure 4). Correlation significance was then tested with an F-test statistic under a linear regression model in R, and additionally verified by using Spearman's Rank Correlation test.

Estimation of the age of node X by using all 72 calibration sets

To achieve an overall estimate for the age of node X (Figure 1), we performed 72 molecular dating analyses, each by using a different calibration set. For each analysis, the distribution of posterior probabilities was used to calculate the mean age of node X, standard deviation and 95% credibility intervals (representing the 2.5^{th} and 97.5^{th} percentile of the posterior distribution; Table 3). The two chronograms derived from the two calibration sets that yielded the oldest and youngest ages for node X, respectively, were overlaid on each other for comparative purposes (Figure 5). To calculate the overall mean for the age of node X, a distribution was drawn from the 1.8×10^5 Markov chain samples generated during the 72 dating analyses by using the statistics package R (R Development Core Team, 2004; Figure 6).

Results

Phylogenetic analyses

The datasets used for the phylogenetic analyses contained a total of 5124 aligned positions or characters (chars) for 74 taxa, comprising three plastid and two nuclear partitions: *rbc*L (1144 chars), *ndh*F (818 chars), *rpl*16-intron (560 chars), nr18S (1634 chars), and nr26S (968 chars; see Table 1).

The optimal models of molecular evolution selected by MrAIC (Nylander, 2005b) were: SYM+I+G for *rbc*L, GTR+G for *ndh*F, GTR+G for *rpl*16-intron, SYM+I+G for nr18S, and GTR+I+G for nr26S. In all cases, the AICc and BIC selection criteria applied in MrAIC (Nylander, 2005b) converged on the same models. Parameter values for these models were estimated simultaneously for each partition during topology and branch length optimization in MrBayes.

The MrBayes majority rule consensus tree with posterior mean branch lengths, clade credibility values, and maximum likelihood bootstrap support values is shown in Figure 1. The tree topology is congruent with published phylogenies (Conti *et al.*, 2002; Rutschmann *et al.*, 2004; Schönenberger and Conti, 2004; Renner, 2004; Sytsma *et al.*, 2004; Wilson *et al.*, 2005). The well-supported Southeast Asian Crypteroniaceae are sister to a clade formed by the Central/South American Alzateaceae and the African Rhynchocalycaceae, Oliniaceae, and Peneaceae. *Memecylon* is weakly supported as sister to Melastomataceae, corroborating the phylogenies of Renner (2004) and Clausing and Renner (2001). Within Melastomataceae, Melastomeae, Miconieae, and Merianieae are well supported as monophyletic. Within Myrtaceae s.l., Psiloxyloideae are sister to the rest of the clade, referred to as Myrtaceae s.s. (syn. Myrtoideae), in agreement with the phylogeny of Wilson *et al.* (2005). Myrteae, Melaleuceae, Eucalypteae, and Leptospermeae are all monophyletic, as in Wilson *et al.* (2005). The position of Vochysiaceae sister to Myrtaceae s.l. confirms the results of Sytsma *et al.* (2004) and Wilson *et al.* (2005). The relationships among the sampled Onagraceae are resolved as in Conti *et al.* (1993), Levin *et al.* (2003), and Berry *et al.* (2004).

Evaluating rate heterogeneity

The LF χ^2 test applied to the phylogram shown in Figure 1 (after removing the outgroups *Duabanga* and *Cuphea*) indicated significant departure from rate constancy: LF χ^2 test value = 1385.8, df = 71, $p = 7.27 \times 10^{-243}$.

Finding the most internally consistent calibration sets by using fossil cross-validation

The average squared deviations (*s* values) calculated in the fossil cross-validation analyses ranged from 0.061 (with calibration set 1) to 0.265 (with calibration sets 50 and 66; see Table 3). The lowest *s* score was generated by the calibration set where all six fossils were assigned to stem nodes, while the highest *s* score was produced by a set with most fossils assigned to crown nodes (Table 3). The use of penalized likelihood (implemented in *r8s*; Sanderson, 2002) instead of Bayesian dating (implemented in *multidivtime*; Kishino *et al.*, 2001; Thorne and Kishino, 2002) to estimate divergence times did not affect the *s* scores in any significant way (data not shown). Therefore, the ranking of the *s* scores remained unchanged.

Relationship between average squared deviation s and nodal distance. The linear regression between the average squared deviations *s* of the 72 calibration sets and the distances in number of nodes between the calibration points of each calibration set produced a multiple R^2 of 0.8967 (Figure 2). The F-test statistic showed a significant correlation (degrees of freedom 1 and 70; $p < 2.2 \times 10^{-16}$). This result was confirmed by Spearman's Rank Correlation test (stem: rho = 0.9561722; $p < 2.2 \times 10^{-16}$). The three calibration sets with the lowest *s* values were 1, 9, and 5 (Figure 2, left), while the three sets with the highest *s* values were 50, 66, and 58 (Figure 2, right). The three calibration sets that deviated the most below the regression line were 18, 10, and 11 (Figure 2, circled), whereas the three sets that deviated the most above the line were 33, 41, and 25 (Figure 2, circled; see Table 3).

The percent deviation of *s* from the regression line with nodal distance ranged from -27.35%, associated with calibration set 18, to +18.45%, associated with calibration set 33 (see Table 3 and Figure 3). Twelve calibration sets had *s* values at least 10% lower than expected based on the regression line with nodal distance (termed "calibration sets A" from now on; see Figure 3, left side). In all 12 sets, fossil 1 was always assigned to node *1a*, and fossil 6 to node *6b,* while the nodal assignments of fossils 3, 4 and 5 varied (see Table 4). More specifically, calibration sets 18, 10, and 11 (Figure 2) showed the *s* values that deviated the most below the regression line (-27.35%; -25.72%, and -23.95%, respectively; Table 3 and Figure 3). Based on these results, we conclude that these three sets are characterized by the highest level of internal consistency corrected for nodal distance (corrected calibration-set consistency). Conversely, 23 calibration sets were associated with *s* values that were at least 10% higher than expected from the regression line (Figure 3, right side; named "calibration sets B" from now on). More specifically, calibration sets 33, 41, and 25 showed the *s* values that deviated the most above the regression line (18.45%; 20.39%; 20.73%, respectively; Table 5 and Figure 3). Based on these results, we conclude that

these three calibration sets are characterized by the lowest level of internal consistency corrected for nodal distance.

To summarize, the calibration sets with the lowest negative values of percent deviation of *s* from the regression line (Figure 2) represent the sets with the highest level of corrected internal consistency among calibration points (Figure 3, left), while the calibration sets with the highest positive values correspond to the sets with the lowest level of corrected consistency (Figure 3, right).

Effect of corrected calibration-set consistency on dating precision. The correlation between the corrected measure of calibration-set consistency (defined as percent deviation of *s* from the regression line with nodal distance) and the percent standard deviation for the age of node X was significant (Figure 4). Linear regression resulted in a multiple R^2 of 0.5975, and the F-test statistic showed a significant correlation (degrees of freedom 1 and 70; $p < 1.788 \times 10^{-16}$). This result was confirmed by Spearman's Rank Correlation test (stem: rho = 0.721429; $p < 2.2 \times 10^{-16}$). The three calibration sets 18, 10, and 11 produced lower standard deviations for the age of node X than the three calibration sets 33, 41, and 25 (Table 3 and Figure 4).

Age of node X estimated by using all 72 calibration sets

The lowest mean age for node X (72.8 mys) was obtained with calibration set 1, the highest (81.46 mys) with calibration set 66 (Table 3). The chronograms corresponding to these calibration sets are shown in Figure 5. As expected, calibration sets with more stem assignments (for ex., 1, 25, and 5) produced younger ages for node X than calibration sets with more crown assignments (e.g., 70, 72, and 66; Table 3), because the deeper a calibration point is placed in a phylogeny, the younger are the age estimates for all other nodes, and vice versa. Consequently, calibration sets with a mixture of stem and crown assignments (e.g., 51 and 61 with three crown and two stem assignments; Table 3) produced intermediate age estimates for node X. The overall mean for the age of node X based on all 72 dating analyses was 77.74 mys, with a standard deviation of 8.51 mys and a 95% credibility interval of 60.89 to 94.21 mys (Figure 6).

Discussion

Despite well-founded theoretical guidelines (Hennig, 1969; Doyle and Donoghue, 1993; Patterson, 1981; Sanderson, 1998; Magallón, 2004), it is often difficult in practice to determine the exact phylogenetic placement of fossils on the basis of their morphological traits (Doyle and Donoghue, 1992; Manchester and Hermsen, 2000). The uncertainty of nodal assignment might be especially severe for the paleobotanical record, due to the generally lower degree of morphological integration in plants as compared to animals (Hennig, 1966; Doyle and Donohue, 1987; Donoghue *et al.*, 1989; Doyle and Donoghue, 1992). While exhaustive cladistic morphological analyses of extinct and extant taxa should allow for more reliable attachment of fossils to specific nodes, such analyses are rarely available (Donoghue *et al.*, 1989; Doyle and Donoghue, 1993; Hermsen *et al.*, 2003). In most cases, practitioners must depend on published descriptions of the morphological features of a fossil, which are often vague and contradicting when it comes to placing it in a phylogeny (Hickey, 1977; Collinson and Pingen, 1992; Pigg *et al.*, 1993; Rozefelds, 1996; Manchester and Hermsen, 2000). Therefore, a careful review of the paleontological literature for a given group of taxa might lead to multiple nodal assignments of a selected fossil that can be equally defended on the basis of morphology.

The possibility of attaching fossils to multiple nodes also depends on the density of taxon sampling in the relevant phylogenetic neighborhood. In many dated chronograms, only one possible assignment exists, because only one taxon was sampled from the pertinent group (e.g., the assignment of fossil Melastomataceae leaves in Conti *et al.*, 2002, or the assignment of fossil Eucalyptoid fruits in Sytsma *et al.*, 2004; see also Sanderson and Doyle, 2001). While unequivocal, this procedure is hardly satisfactory. In the study presented here, we designed taxon sampling in order to allow for multiple nodal assignment possibilities. In fact, five out of the six selected Myrtales fossils (Table 2) could be justifiably assigned to more than one node, based on the morphological traits discussed in the paleobotanical literature (Hickey, 1977; Collinson and Pingen, 1992; Pigg *et al.*, 1993; Rozefelds, 1996). In total, 72 different assignment combinations (calibration sets) were possible, each comprising six calibration points (Table 3).

Finding the most internally consistent calibration sets by using fossil cross-validation

In this study of Crypteroniaceae and related taxa, we use the fossil cross-validation procedure of Near and Sanderson (2004) in a novel way to assess uncertainty in fossil nodal assignment. More specifically, we try to identify the most congruent calibration sets by comparing

the internal consistencies of all 72 calibration sets generated from alternative placements of six fossils (Table 3). The procedure revealed large differences among the average squared deviation *s* associated with the 72 calibration sets, ranging from an *s* value of 0.061 for set 1 to an *s* value of 0.265 for set 66 (Table 3; Figure 2). Therefore, one might conclude that the assignment of fossils 1, 2, 3, 4, 5, and 6 to nodes *1a, 2, 3a, 4a, 5a*, and *6a*, respectively, produced the most internally consistent calibration set, while the assignment of the same fossils to nodes *1c, 2, 3c, 4b, 5b,* and *6b*, respectively, the most inconsistent one. However, in calibration set 1 all the fossils are assigned to their stem nodes, while in calibration set 66 all the fossils are assigned to their crown nodes, except for fossil 2, for which only one nodal assignment is possible (Table 3). It is thus reasonable to ask whether the *s* values might be influenced by the nodal distance among calibration points. Indeed, a significant positive correlation between the average squared deviation *s* and nodal distances was observed for the calibration sets (Figure 2).

The positive correlation between *s* values and nodal distances found in our study differs from the results described in Near *et al.* (2005a). In their study of centrarchid fishes, the authors plotted the percent deviation between molecular and fossil ages in a single calibration set versus the number of nodes separating all possible pairs of calibration points in the set. Their results showed no correlation between percent deviations and nodal distances, leading them to conclude that the proximity among calibration points in a phylogeny has no effect on the results of fossil cross-validation. The key differences between their and our results may depend in part on methodological details (i.e., the different procedures used to calculate nodal distances in the two studies), but also on the fact that, as Near *et al.* (2005a) remarked, their fossils were more or less evenly distributed across the phylogeny, while in our case fossils are concentrated in three clades (Melastomataceae, Myrtaceae s.l., Onagraceae; Figure 1).

How can we then explain the correlation between *s* values and nodal distance found in our study (Figure 2)? It is important to remember that the *s* values are essentially derived from the difference between the estimated and the fossil ages of the calibration points in a set. Therefore, one possible interpretation of the observed correlation might relate to the procedures used for nodal age estimation. More specifically, the rate smoothing methods employed in our dating analyses, based on both Bayesian (Thorne *et al.*, 1998) and penalized likelihood (Sanderson, 2002) approaches, allow rates to change between ancestral-descendant branches, thus creating estimation errors that depend on the number of nodes involved in the smoothing procedure. Consequently, the greater the nodal distance among calibration points, the greater the possible difference between the estimated and the fossil ages of the calibration points in a set. Because greater nodal distances are associated with sets where fossils are mostly assigned to the corresponding crown nodes (see Figure 1 and Table 3), this would explain why such sets are characterized by greater *s* values, as in the case of set

66 (Figure 2). Conversely, calibration sets where most fossils are attached to the stem nodes, as in the case of set 1, are associated with smaller nodal distances among calibration points, hence with smaller *s* values (Figure 2; Table 3). Thus, our results suggest that the smaller *s* values associated with calibration sets where most fossils are assigned to stem nodes do not inherently reflect a higher level of consistency among the calibration points, but the effects of nodal distance. Therefore, *s* values appear to represent a biased estimate of internal consistency.

In order to identify the calibration sets least and most affected by nodal distance bias, we ranked all sets according to the percent deviation of *s* from the regression line with nodal distance (Figure 3). This allowed us to recognize calibrations sets 18, 10 and 11 as those associated with the highest level of internal consistency corrected for nodal distance, and sets 33, 41, and 25 as those with the lowest level of corrected internal consistency (Table 5). Importantly, the most consistent calibration sets also produced the lowest percent standard deviations for the estimated ages of node X, and the least consistent sets the highest percent standard deviations (Figure 4). This positive correlation might be explained by the observation that inconsistent calibration points in a set contradict each other in their statements about the timing of evolutionary events, thus producing conflicting estimates for the age(s) of the node(s) of interest, hence higher associated errors (Near and Sanderson, 2004; Near *et al.*, 2005b). Conversely, calibration points in sets with high levels of corrected internal consistency produce convergent estimates for the age(s) of the node(s) of interest, hence lower associated errors.

The 12 calibration sets associated with the highest level of corrected internal consistency (Figure 3, left side) share some common properties. In all, the temporal information provided by fossil 1 is most consistent with that of the other calibration points if it is assigned to node *1a* (Table 4 and Figure 1). This result supports the interpretation by Renner *et al.* (2001) and Sytsma *et al.* (2004) that the fossil leaves from the Early Eocene of North Dakota (Hickey, 1977) should be assigned to the node representing the entire Melastomataceae crown group, because of the acrodromous leaf venation. On the other hand, fossil 6, representing the pollen *Diporites aspis* from the Early Oligocene of Otway Basin (Australia; Berry *et al.*, 1990; Table 2), is most consistent with the other calibration points if it is assigned to node *6b*, representing the *Fuchsia* crown group (Table 4 and Figure1). This conclusion is congruent with the results of molecular dating analyses that produced an age interval for the node corresponding to *6b* compatible with the age of *Diporites aspis* (Berry *et al.*, 2004; Sytsma *et al.*, 2004).

No clear pattern emerges from comparisons among calibration sets A for the assignment of fossils 3, 4, and 5 (Table 4). However, in the calibration set with the highest level of corrected internal consistency (18; see Table 5), the three fossils are all assigned to their crown positions (nodes *3c*, *4b*, and *5b*; Figure 1). Based on this evidence, the pollen *Myrtaceidites lisamae* from the

Santonian of Gabon (fossil 3) is most consistently attributed to Myrtaceae s.s. (node *3c*), as proposed by Sytsma *et al.* (2004). Also in agreement with Sytsma *et al.* (2004), the fruits and seeds of *Paleomyrtinaea* from the latest Paleocene of North Dakota (fossil 4) are most consistently placed with the crown radiation of the tribe Myrteae (Myrtoideae s.s.; node *4b*). Finally, the Eucalypt-like fruits from the Middle Eocene of South Eastern Queensland (fossil 5) are best assigned to the node representing the most recent common ancestor of *Eucalyptus* and *Angophora* (node *5b*), as suggested by Rozefelds (1996).

Age of node X estimated by using all 72 calibration sets

The overall mean age for node X estimated from analyses performed with all 72 calibration sets was 77.74 mys, with a credibility interval of 60.89 to 94.21 mys (see Figure 6). By using all possible calibration sets in the dating analyses, all known uncertainties of nodal assignment are incorporated in the final age estimate for the node of interest, providing increased confidence that the "real" age of the node is contained in the estimate. It is worth noticing that the ages of node X (79.70 mys; 79.14 mys, and 78.15 mys; Table 5) estimated from the three calibration sets (18, 10, 11; respectively) with the highest level of corrected internal consistency fall within the credibility interval of the overall mean age. Therefore, it might be suggested that the "overall mean approach" increases the accuracy of the estimate, lending further support to the proponents of multi-calibration (Lee, 1999; Conroy and van Tuinen, 2003; Reisz and Müller, 2004).

At the same time, however, the "overall mean approach" reduces the precision of the estimate. In fact, the overall mean age is associated with a higher standard deviation (8.51 mys) than those of the ages estimated by using the three calibration sets with the highest corrected calibration-set consistency (18: 7.55 mys; 10: 7.34 mys; 11: 7.49 mys; Table 5). This observation might again be explained by the fact that all the biases introduced by conflicting nodal assignments of the fossils in the "overall mean approach" contribute to the errors associated with the estimated ages (Conroy and van Tuinen, 2003; Conti *et al.*, 2004; Reisz and Müller, 2004; Müller and Reisz, 2005).

All estimates generated in this study (Table 3) for the mean age of the split between the Southeast Asian Crypteroniaceae and their West Gondwanan sister clade are contained within the interval ranging from 81.5 to 72.8 mys (Figure 5). This range overlaps almost entirely with the results of published age estimates for the same node (62 to 109 mys, Rutschmann *et al.*, 2004; 57 to 79 mys, Moyle, 2004), even though the mentioned studies differed from ours both in gene/taxon sampling and calibration strategies. Despite these differences, the biogeographic scenario most compatible with the timeframes calculated for node X is that the Crypteroniaceae stem lineage dispersed from Africa to the Deccan plate as it drifted northward during the Late Cretaceous

(approximately 125 to 84 mys ago; Plummer and Belle, 1995; McLoughlin, 2001). The newly obtained age estimates, then, further support India's likely role in expanding the range of Crypteroniaceae from Africa to Asia during its northbound movement along the African coast, corroborating the out-of-India hypothesis for the origin of Crypteroniaceae (Conti *et al.*, 2002; Conti *et al.*, 2004; Moyle, 2004; Rutschmann *et al.*, 2004; see also Lieberman, 2003).

To summarize, our study illustrates two main approaches to the problem of nodal fossil assignment when equally justifiable alternatives exist. On the one hand, one might use a modified version of the fossil cross-validation procedure to identify the calibration sets with the highest level of internal consistency, corrected for nodal distance bias, and then use these sets to estimate the ages of the nodes of interest. Such sets should generate lower standard deviations associated with nodal age estimates than sets characterized by lower levels of correct consistency. On the other hand, estimating the overall mean ages for the nodes of interest by using all possible calibration sets might represent a more cautious approach.

While we have attempted to suggest practical procedures, based on available methodology (i.e., fossil cross-validation; Near and Sanderson, 2004; Near *et al.*, 2005b), to address the difficult problem of fossil nodal assignment, we also wish to emphasize that such measures can by no means replace careful review, selection, and evaluation of the fossil record used for calibration. To further improve confidence in the assignment of selected fossils to specific nodes in a phylogeny, a multi-pronged approach will be necessary, including comprehensive morphological cladistic analyses of extinct and extant taxa (Donoghue *et al.*, 1989; Doyle, 2000; Eklund *et al.*, 2004), quantitative evaluation of the gap between the time of lineage divergence and the time of first appearance of synapomorphies in the fossil record (Foote and Sepkoski, 1999; Tavaré *et al.*, 2002), improved paleontological dating of fossils, and evaluation of the positional effects of calibration nodes in relation to the nodes of interest (Smith and Peterson, 2002; Conroy and van Tuinen, 2003; Porter *et al.*, 2005).

References

Alroy, J. 1999. The fossil record of North American mammals: evidence for a Paleocene evolutionary radiation. *Systematic Biology* **48**:107-118.

Aris-Brosou, S., and Z. Yang. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Systematic Biology* **51**:703-714.

Ashton, P. S., and C. V. S. Gunatilleke. 1987. New light on the plant geography of Ceylon: I. Historical plant geography. *Journal of Biogeography* **14**:249-285.

Baum, D. A., R. L. Small, and J. F. Wendel. 1998. Biogeography and floral evolution of Baobabs (*Adansonia*, Bombacaceae) as inferred from multiple data sets. *Systematic Biology* **47**:181-207.

Bell, C. D., and M. J. Donoghue. 2005. Dating the Dipsacales: Comparing models, genes, and evolutionary implications. *American Journal of Botany* **92**:284-296.

Benton, M. J., and F. J. Ayala. 2003. Dating the tree of life. *Science* **300**:1698-1700.

Berry, P. E., W. J. Hahn, K. J. Sytsma, J. C. Hall, and A. Mast. 2004. Phylogenetic relationships and biogeography of *Fuchsia* (Onagraceae) based on noncoding nuclear and chloroplast DNA data. *American Journal of Botany* **94**:601-614.

Berry, P. E., J. J. Skvarla, A. D. Partridge, and M. K. Macphail. 1990. *Fuchsia* pollen from the Tertiary of Australia. *Australian Systematic Botany* **3**:739-744.

Boltenhagen, E. 1976. Pollens et spores Sénoniennes du Gabon. *Cahiers de Micropaléontologie* **17**:1-21.

Bossuyt, F., and M. C. Milinkovitch. 2001. Amphibians as indicators of early Tertiary "out-of-India" dispersal of Vertebrates. *Science* **292**:93-95.

Brochu, C. A. 2004. Calibration age and quartet divergence date estimation. *Evolution* **58**:1375-1382.

Bult, C., M. Källersjö, and Y. Suh. 1992. Amplification and sequencing of 16/18S rDNA from gel-purified total plant DNA. *Plant Molecular Biology Reporter* **10**:273-284.

Christophel, D. C., L. J. Scriven, and D. R. Greenwood. 1992. An Eocene megafossil flora from Nelly Creek, South Australia. *Transactions of the Royal Society of South Australia* **116**:65-76.

Clausing, G., and S. S. Renner. 2001. Molecular phylogenetics of Melastomataceae and Memecylaceae: implications for character evolution. *American Journal of Botany* **88**:486-498.

Collinson, M. E., and M. Pingen. 1992. Seeds of the Melastomataceae from the Miocene of Central Europe. Pp. 129-139 *in* J. Kovar-Eder, ed. Palaeovegetational development in Europe. Museum of Natural History, Vienna, Austria.

Conroy, C. J., and M. van Tuinen. 2003. Extracting time from phylogenies: positive interplay between fossil and genetic data. *Journal of Mammalogy* **84**:444-455.

Conti, E., T. Eriksson, J. Schönenberger, K. J. Sytsma, and D. A. Baum. 2002. Early Tertiary out-of-India dispersal of Crypteroniaceae: evidence from phylogeny and molecular dating. *Evolution* **56**:1931-1942.

Conti, E., A. Fischbach, and K. J. Sytsma. 1993. Tribal relationships in Onagraceae: implications from *rbc*L sequence data. *Annals of the Missouri Botanical Garden* **80**:672-685.

Conti, E., A. Litt, and K. J. Sytsma. 1996. Circumscription of Myrtales and their relationsships to other Rosids: evidence from *rbc*L sequence data. *American Journal of Botany* **83**:221-233.

Conti, E., A. Litt, P. G. Wilson, S. A. Graham, B. G. Briggs, L. A. S. Johnson, and K. J. Sytsma. 1997. Interfamilial relationships in Myrtales: molecular phylogeny and patterns of morphological Evolution. *Systematic Botany* **22**:629-647.

Conti, E., F. Rutschmann, T. Eriksson, K. J. Sytsma, and D. A. Baum. 2004. Calibration of molecular clocks and the biogeographic history of Crypteroniaceae: A reply to Moyle. *Evolution* **58**:1874-1876.

Cowles, M. K., and B. P. Carlin. 1996. Markov chain Monte Carlo convergence diagnosis: A comparative review. *Journal of the American Statistical Association* **91**:883-904.

Crane, P. R., S. R. Manchester, and D. L. Dilcher. 1990. A preliminary survey of fossil leaves and well-preserved reproductive stuctures from the Sentinel Butte Formation (Paleocene) near Almont, North Dakota. *Fieldiana Geology* **1418**:1-63.

Daghlian, C. P., J. J. Skvarla, D. T. Pocknall, and P. H. Raven. 1985. *Fuchsia* pollen from the early Miocene of New Zealand. *American Journal of Botany* **72**:1039-1047.

Dahlgren, R., and R. F. Thorne. 1984. The order Myrtales: circumscription, variation, and relationships. *Annals of the Missouri Botanical Garden* **71**:633-699.

Dahlgren, R., and A. E. Van Wyk. 1988. Structures and relationships of families endemic to or centered in Southern Africa. *Monographs in Systematic Botany from the Missouri Botanical Garden* **25**:1-94.

Darwin, C. 1859. On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life. John Murray, London.

de Queiroz, K., and J. Gauthier. 1990. Phylogeny as a central principle in taxonomy: phylogenetic definitions of taxon names. *Systematic Zoology* **39**:307-322.

Donoghue, M. J., J. A. Doyle, J. Gauthier, A. G. Kluge, and T. Rowe. 1989. The importance of fossils in phylogeny reconstruction. *Annual Reviews of Ecology and Systematics* **20**:431-460.

Dorofeev, P. I. 1960. On the Tertiary flora of Belorussia. *Botanichesky Zhurnal SSSR* **45**:1418-1434 (in Russian).

Dorofeev, P. I. 1963. The Tertiary floras of western Siberia. Izdatelstvo Akademii Nauk SSSR. Moskva-Leningrad, Russia (in Russian).

Dorofeev, P. I. 1988. Miocene floras of the Tambov district. Akademii Nauk, Leningrad, Russia (in Russian, posthumous work, F.Y. Velichkevich, ed.).

Doyle, J. A. 2000. Paleobotany, relationships, and geographic history of Winteraceae. *Annals of the Missouri Botanical Garden* **87**:303-316.

Doyle, J. A., and M. J. Donoghue. 1987. The importance of fossils in elucidating seed plant phylogeny and macroevolution. *Review of Paleobotany and Palynology* **50**:63-95.

Doyle, J. A., and M. J. Donoghue. 1992. Fossils and seed plant phylogeny reanalyzed. *Brittonia* **44**:89-106.

Doyle, J. A., and M. J. Donoghue. 1993. Phylogenies and angiosperm diversification. *Paleobiology* **19**:141-167.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**:e88.

Dyjor, S., Z. Kvacek, M. Lanucka-Srodoniowa, W. Pyszynski, A. Sadowska, and E. Zastawniak. 1992. The Younger Tertiary deposits in the Gozdnica region (SW Poland) in the light of recent palaeobotanical research. *Polish Botanical Studies* **3**:1-129.

Eklund, H., J. A. Doyle, and P. S. Herendeen. 2004. Morphological phylogenetic analysis of living and fossil Chloranthaceae. *International Journal of Plant Sciences* **165**:107-151.

Fairon-Demaret, M. 1994 [1996]. Les fruits et graines du Miocene de Bioul (Entre-Samre-et-Meuse, Belgique). Etude qualitative, quantitative et considerations paleoécologiques. *Annales de la Societé Géologique de Belgique* **117**:277-309.

Felsenstein, J. 2004. PHYLIP (Phylogeny Inference Package), version 3.63. Department of Genetics, University of Washington, Seattle, WA.

Foote, M., J. P. Hunter, C. M. Janis, and J. J. Sepkoski Jr. 1999. Evolutionary and preservational constraints on origins of biologic groups: divergence times of Eutherian mammals. *Science* **283**:1310-1314.

Foote, M., and J. J. Sepkoski Jr. 1999. Absolute measures of the completeness of the fossil record. *Nature* **398**:415-417.

Graham, S. A. 1984. Alzateaceae, a new family of Myrtales in the American Tropics. *Annals of the Missouri Botanical Garden* **71**:757-779.

Graur, D., and W. Martin. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *TRENDS in Genetics* **20**:80-86.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**:696-704.

Hedges, S. B., and S. Kumar. 2003. Genomic clocks and evolutionary timescales. *TRENDS in Genetics* **19**:200-206.

Hennig, W. 1966. Phylogenetic Systematics. University Illinois Press, Urbana, Illinois.

Hennig, W. 1969. Die Stammesgeschichte der Insekten. Kramer, Frankfurt, Germany.

Hermsen, E. J., M. A. Gandolfo, K. C. Nixon, and W. L. Crepet. 2003. *Divisestylus* gen. nov. (aff. Iteaceae), a fossil saxifrage from the Late Cretaceous of New Jersey, USA. *American Journal of Botany* **90**:1373-1388.

Herngreen, G. F. W. 1975. An Upper Senonian pollen assemblage of borehole 3-PIA-10-AL, State of Alagoas, Brazil. *Pollen Spores* **17**:93-140.

Hickey, L. J. 1977. Stratigraphy and paleobotany of the Golden Valley formation (Early Tertiary) of western North Dakota. Memoir 150. Geological Society of America, Boulder, Colorado.

Ho, S. Y. W., and G. Larson. 2006. Molecular clocks: when the times are a-changin'. *TRENDS in Genetics* **22**:79-83.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* **17**:754.

Johnson, L. A. S., and B. G. Briggs. 1984. Myrtales and Myrtaceae - a phylogenetic analysis. *Annals of the Missouri Botanical Garden* **71**:700-756.

Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* **18**:352-361.

Kuzoff, R. K., J. A. Sweere, D. E. Soltis, P. S. Soltis, and E. A. Zimmer. 1998. The phylogenetic potential of entire 26S rDNA sequences in plants. *Molecular Biology and Evolution* **15**:251-263.

Langley, C. H., and W. Fitch. 1974. An estimation of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution* **3**:161-177.

Lantern, L. R., and W. P. Sharp. 1989. Seed coat characters of some American Myrtinae (Myrtaceae): *Psidium* and related genera. *Systematic Botany* **14**:370-376.

Lantern, L. R., and D. Stevenson. 1986. Variability of embryos in subtribe Myrtinae (Myrtaceae). *Systematic Botany* **11**:155-162.

Lee, M. S. Y. 1999. Molecular clock calibrations and Metazoan divergence dates. *Journal of Molecular Evolution* **49**:385-391.

Levin, R. A., W. L. Wagner, P. C. Hoch, M. Nepokroff, J. C. Pires, E. A. Zimmer, and K. J. Sytsma. 2003. Family-level relationships of Onagraceae based on chloroplast *rbc*L and *ndh*F data. *American Journal of Botany* **90**:107-115.

Lieberman, B. S. 2003. Unifying theory and methodology in biogegraphy. *Journal of Evolutionary Biology* **33**:1-25.

Macey, J. R., J. A. Schulte II, A. Larson, N. B. Ananjeva, Y. Wang, R. Rethiyagoda, N. Rastegar-Pouyani, and T. J. Papenfuss. 2000. Evaluating trans-Tethys migration: an example using acrodont lizard phylogenetics. *Systematic Biology* **49**:233-256.

Maddison, P. G., and D. R. Maddison. 2000. MacClade 4: analysis of phylogeny and character evolution. Sinauer, Sunderland, MA.

Magallón, S. A. 2004. Dating lineages: molecular and paleontological approaches to the temporal framework of clades. *International Journal of Plant Sciences* **165**:7-21.

Magallón, S. A., and M. J. Sanderson. 2001. Absolute diversification rates in Angiosperm clades. *Evolution* **55**:1762-1780.

Magallón, S. A., and M. J. Sanderson. 2005. Angiosperm divergence times: the effect of genes, codon positions, and time constraints. *Evolution* **59**:1653-1670.

Mai, D. H. 1995. Tertiäre Vegetationsgeschichte Europas. G. Fischer, Jena, Germany.

Mai, D. H. 2000. Die untermiozänen Floren aus der Spremberger Folge und dem II. Flözhorizont der Lausitz. Teil III. Dialypetalae und Sympetalae. *Palaeontographica Abteilung B* **253**:1-106.

Manchester, S. R. 1999. Biogeographical relationships of North American Tertiary floras. *Annals of the Missouri Botanical Garden* **86**:472-522.

Manchester, S. R., and E. J. Hermsen. 2000. Flowers, fruits, seeds, and pollen of *Landeenia* gen. Nov., an extinct sapindalean genus from the Eocene of Wyoming. *American Journal of Botany* **87**:1909-1914.

Marshall, C. R. 1990a. Confidence intervals on stratigraphic ranges. *Paleobiology* **16**:1-10.

Marshall, C. R. 1990b. The fossil record and estimating divergence times between lineages - maximum divergence times and the importance of relieable phylogenies. *Journal of Molecular Evolution* **30**:400-408.

McKenna, M. C. C. 1973. Sweepstakes, filters, corridors, Noah's arks, and beached Viking funeral ships in paleogeography. Pp. 291-304 *in* D. H. Tarling and S. K. Runcorn, eds. Implications of Continental Drift to the Earth Sciences. Academic Press, London.

McLaughlin, S. 2001. The breakup history of Gondwana and its impact on pre-Cenozoic floristic provincialism. *Australian Journal of Botany* **48**:271-300.

Mentink, H., and P. Baas. 1992. Leaf anatomy of the Melastomataceae, Memecylaceae, and Crypteroniaceae. *Blumea* **37**:189-225.

Morley, R. J. 2000. Origin and evolution of tropical rain forests. John Wiley & Sons Ltd., Chichester, England.

Morley, R. J., and C. W. Dick. 2003. Missing fossils, molecular clocks and the origin of the Melastomataceae. *American Journal of Botany* **90**:1638-1645.

Moyle, R. G. 2004. Calibration of molecular clocks and the biogeographic history of Crypteroniaceae. *Evolution* **58**:1871-1873.

Muller, J. 1968. Palynology of the Pedawan and Plateau Sandstone Formations (Creataceous-Eocene) in Sarawak, Malaysia. *Micropaleontology* **14**:1-37.

Muller, J. 1981. Fossil pollen records of extant angiosperms. *Botanical Reviews* **47**:1-142.

Müller, J., and R. R. Reisz. 2005. Four well-constrained calibration points from the vertebrate fossil record for molecular clock estimates. *BioEssays* **27**:1069-1075.

Near, T. J., D. I. Bolnick, and P. C. Wainwright. 2005a. Fossil calibrations and molecular divergence time estimates in centrarchid fishes (Teleostei: Centrarchidae). *Evolution* **59**:1768-1782.

Near, T. J., P. A. Meylan, and H. B. Shaffer. 2005b. Assessing concordance of fossil calibration points in molecular clock studies: an example using turtles. *American Naturalist* **165**:137-146.

Near, T. J., and M. J. Sanderson. 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **359**:1477-1483.

Nylander, J. 2005a. BootPHYML 3.4. School of Computational Science (SCS), Florida State University, Tallahassee, Florida.

Nylander, J. 2005b. MrAIC 1.4. School of Computational Science (SCS), Florida State University, Tallahassee, Florida.

Olmstead, R. G., and J. A. Sweere. 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. *Systematic Biology* **43**:467-481.

Patterson, C. 1981. Significance of fossils in determining evolutionary relationships. *Annual Reviews of Ecology and Systematics* **12**:195-223.

Penny, D. 2005. Relativity for molecular clocks. *Nature* **436**:183-184.

Pereira, J. T. 1996. Crypteroniaceae. Pp. 135-149 *in* E. Soepadmo, K. M. Wong and L. G. Saw, eds. Tree flora of Sabah and Sarawak, Volume 2. Forest Research Institute Malaysia, Sabah Forestry Department, and Sarawak Forestry Department, Kuala Lumpur, Malaysia.

Pereira, J. T., and K. M. Wong. 1995. Three new species of *Crypteronia* (Crypteroniaceae) from Borneo. *Sandakania* **6**:41-53.

Pigg, K. B., R. A. Stockey, and S. L. Maxwell. 1993. *Paleomyrtinaea*, a new genus of permineralized myrtaceous fruits and seeds from the Eocene of British Columbia and Paleocene of North Dakota. *Canadian Journal of Botany* **71**:1-9.

Plummer, P. S., and E. R. Belle. 1995. Mesozoic tectono-stratigraphic evolution of the Seychelles microcontinent. *Sedimentary Geology* **96**:73-91.

Porter, M. L., M. Pérez-Losada, and K. A. Crandall. 2005. Model-based multi-locus estimation of decapod phylogeny and divergence times. *Molecular Phylogenetics and Evolution* **37**:355-369.

R Development Core Team. 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org.

Rambaut, A., and A. J. Drummond. 2005. Tracer 1.3. A program for analyzing results from Bayesian MCMC programs such as BEAST & MrBayes. Department of Zoology, University of Oxford, Oxford, UK. Available from http://evolve.zoo.ox.ac.uk/software.html.

Reisz, R. R., and J. Müller. 2004. Molecular timescales and the fossil record: a paleontological perspective. *TRENDS in Genetics* **20**:237-241.

Renner, S. S. 2004. Bayesian analysis of combined chloroplast loci, using multiple calibrations, supports the recent arrival of Melastomataceae in Africa and Madagascar. *American Journal of Botany* **91**:1427-1435.

Renner, S. S. 2005. Relaxed molecular clocks for dating historical plant dispersal events. *TRENDS in Plant Science* **10**:550-558.

Renner, S. S., G. Clausing, and K. Meyer. 2001. Historical biogeography of Melastomataceae: the roles of Tertiary migration and long-distance dispersal. *American Journal of Botany* **88**:1290-1300.

Renner, S. S., and K. Meyer. 2001. Melastomeae come full circle: biogeographic reconstruction and molecular clock dating. *Evolution* **55**:1315-1324.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**:1572-1574.

Rozefelds, A. C. 1996. *Eucalyptus* phylogeny and history: a brief summary. *Tasforests* **8**:15-26.

Rutschmann, F. 2005. Bayesian molecular dating using paml/multidivtime. A step-by-step manual. Version 1.5 (July 2005). Institute of Systematic Botany, University of Zurich, Zurich, Switzerland. Available from http://www.plant.ch.

Rutschmann, F. 2006. Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Diversity and Distributions* **12**:35-48.

Rutschmann, F., T. Eriksson, J. Schönenberger, and E. Conti. 2004. Did Crypteroniaceae really disperse out of India? Molecular dating evidence from *rbc*L, *ndh*F, and *rpl*16 sequences. *International Journal of Plant Sciences* **165**:69-83.

Sanderson, M. J. 1997. A nonparameteric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* **14**:1218-1231.

Sanderson, M. J. 1998. Estimating rate and time in molecular phylogenies: beyond the molecular clock? Pp. 242-264 *in* D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. Molecular systematics of plants II; DNA sequencing. Kluwer Academic Publishers, Norwell, MA.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a Penalized Likelihood approach. *Molecular Biology and Evolution* **19**:101-109.

Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**:301-302.

Sanderson, M. J., J. L. Thorne, N. Wikström, and K. Bremer. 2004. Molecular evidence on plant divergence times. *American Journal of Botany* **91**:1656-1665.

Schönenberger, J., and E. Conti. 2003. Molecular phylogeny and floral evolution of Penaeaceae, Oliniaceae, Rhynchocalycaceae, and Alzateaceae (Myrtales). *American Journal of Botany* **90**:293-309.

Sebola, R. J., and K. Balkwill. 1999. Resurrection of two previously confused species, *Olinia capensis* (Jacq.) Klotzsch and *O. micrantha* Decne. (Oliniaceae). *South African Journal of Botany* **65**:97-103.

Smith, A. B., and J. J. Peterson. 2002. Dating the time of origin of major clades: molecular clocks and the fossil record. *Annual Review of Earth and Planetary Sciences* **30**:65-88.

Soltis, P. S., D. E. Soltis, V. Savolainen, P. R. Crane, and T. G. Barraclough. 2002. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and evidence for molecular living fossils. *Proceedings of the National Academy of Sciences USA* **99**:4430-4435.

Springer, M. S. 1995. Molecular clocks and the incompleteness of the fossil record. *Journal of Molecular Evolution* **41**:531-538.

Storey, M., J. J. Mahoney, A. D. Saunders, R. A. Duncan, S. P. Kelley, and M. F. Coffin. 1995. Timing of hot spot-related volcanism and the breakup of Madagascar and India. *Science* **267**:852-855.

Swofford, D. L. 2001. PAUP* 4.0b10: Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer, Sunderland, MA.

Sytsma, K. J., A. Litt, M. L. Zjhra, J. C. Pires, M. Nepokroeff, E. Conti, J. Walker, and P. G. Wilson. 2004. Clades, clocks, and continents: historical and biogeographical analysis of Myrtaceae, Vochysiaceae, and relatives in the southern hemisphere. *International Journal of Plant Sciences* **165**:85-105.

Tavaré, S., C. R. Marshall, O. Will, C. Soligo, and R. D. Martin. 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature* **416**:726-729.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **24**:4876-4882.

Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**:689-702.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**:1647-1657.

Tobe, H., and P. H. Raven. 1984. The embryology and relationships of Oliniaceae. *Plant Systematics and Evolution* **146**:105-116.

van der Hammen, T. 1954. El desarrollo de la flora colombiana en los periodos geológicos. I. Maestrichtiano hasta Terciario más inferior. *Boletín de Geología* **2**:49-106.

van Tuinen, M., and E. A. Hadly. 2004. Error in estimation of rate and time inferred for the early Amniote fossil record and Avian molecular clocks. *Journal of Molecular Evolution* **59**:267-276.

Welch, J. J., and L. Bromham. 2005. Molecular dating when rates vary. *TRENDS in Ecology and Evolution* **20**:320-327.

Wikström, N., V. Savolainen, and M. W. Chase. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society B: Biological Sciences* **268**:2211-2220.

Wikström, N., V. Savolainen, and M. W. Chase. 2003. Angiosperm divergence times: congruence and incongruence between fossils and sequence divergence estimates. Pp. 142-165 *in* P. C. J. Donoghue and M. P. Smith, eds. Telling the evolutionary time: molecular clocks and the fossil record. Taylor & Francis, London, UK.

Wilson, P. G., M. M. O'Brien, M. M. Heslewood, and C. J. Quinn. 2005. Relationships within Myrtaceae sensu lato based on a *mat*K phylogeny. *Plant Systematics and Evolution* **251**:3-19.

Wray, G. A. 2001. Dating branches on the Tree of Life using DNA. *Genome Biology* **3**:1.1-1.7.

Yang, Z. 2004. A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times. *Acta Zoologica Sinica* **50**:645-656.

Yang, Z., and B. Rannala. 2005. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution* **23**:212-226.

Zuckerkandl, E., and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. Pp. 97-166 *in* V. Bryson and H. Vogel, eds. Evolving Genes and Proteins. Academic Press, New York.

Zurawski, G., B. Perrot, W. Bottomley, and P. R. Whitfield. 1981. The structure of the gene for the large subunit of ribulose-1,5-bisphosphate carboxylase from spinach chloroplast DNA. *Nucleic Acids Research* **14**:3251-3270.

Tables

Table 1. Species names, sources and GenBank accession numbers for the DNA sequences used in this study. Accession numbers with an asterisk represent newly generated and submitted sequences. The other sequences were from [a]Schönenberger and Conti, 2003, [b]Conti *et al.*, 1996, [c]Clausing and Renner, 2001, [d]Conti *et al.*, 2002, [e]Renner *et al.*, 2001, [f]Renner and Meyer, 2001, [g]Rutschmann *et al.*, 2004. n.a. = sequences could not be generated and were treated as missing data. Herbaria acronyms: BOL = Bolus, University of Cape Town, South Africa; BRUN = Brunei Forestry Centre, Brunei-Darussalam; CANB = Centre for Plant Biodiversity Research, Canberra, Australia; CAY = Institut de Recherche pour le Developpement, French Guiana/Cayenne, INPA = Instituto Nacional de Pesquisas da Amazônia, Manaus, Brazil; K = Royal Botanic Gardens, Kew, England; NY = New York Botanical Garden, Bronx, New York; SAN = Forest Research Centre, Sandakan, Sabah, Malaysia; Z = University of Zurich, Switzerland.

Taxon	**Voucher**	**GenBank accession numbers**				
		***rbc*L**	***ndh*F**	***rpl*16-intron**	**nr18S**	**nr26S**
Crypteroniaceae A. DC. (1868), nom. cons.						
Axinandra coriacea Baill.	J. Pereira, san 142218-142220, Sabah, Malaysia, (SAN), (Z).	AM235621	n.a.	AM235441	AM235477	AM235549
Axinandra zeylanica Thwaites	Peter Ashton, s.n., Sri Lanka	AY078157[d]	AJ605094[g]	AJ605107[g]	AM235478	AM235550
Crypteronia borneensis J.T. Pereira & Wong	F. Rutschmann, fru 61, (BRUN), (Z)	AM235622	AM235389	AM235442	AM235479	AM235551
Crypteronia glabrifolia J.T. Pereira & Wong	F. Rutschmann, fru 28, (BRUN), (Z)	AM235623	AM235390	AM235443	AM235480	AM235552
Crypteronia griffithii C. B. Clarke	Shawn Lum, s.n., Singapore	AJ605087[g]	AJ605098[g]	AJ605108 [g]	AM235481	AM235553
Crypteronia paniculata Blume	See GenBank associated references	AY078153[d]	AJ605099[g]	AY151597[a]	AM235482	AM235554

Dactylocladus stenostachys Oliver	Peter Becker, s.n., Brunei	AY078156[d]	AJ605100[g]	AJ605109[g]	AM235483	AM235555
Alzateaceae S.A. Graham (1985)						
Alzatea verticillata Ruiz & Pavon	See GenBank associated references	U26316[b]	AF215591[c]	AY151598[a]	AM235484	AM235556
Rhynchocalycaceae L. A.S. Johnson & B.G. Briggs (1985)						
Rhynchocalyx lawsonioides Oliver	See GenBank associated references	U26336[b]	AF270757[c]	AY151599[a]	AM235485	AM235557
Oliniaceae Arn. ex Sond. (1839), nom. cons.						
Olinia capensis (Jacq.) Klotzsch	J. Schönenberger 519 (Z), (BOL)	AM235624	AM235392	AY151600[a]	AM235486	AM235558
Olinia emarginata Burtt Davy	J. Schönenberger 579, cultivated in Kirstenbosch Botanical Garden, (Z)	AJ605089[g]	AJ605102[g]	AY151601[a]	AM235487	AM235559
Olinia radiata Hofmeyr & Phill.	T. Abbott, 6341, (Z)	AM235625	AM235393	AY151602[a]	AM235488	AM235560
Olinia vanguerioides Baker f.	A. Blarer, s.n., (Z)	AM235626	AM235394	AY151603[a]	AM235489	AM235561
Olinia ventosa (L.) Cufod.	See GenBank associated references	AF215546[c]	AF215594[c]	AY151604[a]	AM235490	AM235562
Penaeaceae Sweet ex. Guillemin (1828), nom. cons.						
Brachysiphon acutus (Thunb.) A. Juss.	J. Schönenberger 365, (Z), (BOL)	AJ605084[g]	AJ605095[g]	AY151605[a]	AM235491	AM235563
Brachysiphon fucatus (L.) Gilg	J. Schönenberger 357 (Z), (BOL)	AJ605085[g]	AJ605096[g]	AY151606[a]	AM235492	AM235564
Brachysiphon microphyllus Rourke	J. Schönenberger 386 (Z), (BOL)	AJ605086[g]	AJ605097[g]	AY151608[a]	AM235493	AM235565
Brachysiphon mundii Sond.	J. Schönenberger 377 (Z), (BOL)	AM235627	AM235395	AY151607[a]	AM235494	AM235566

Brachysiphon rupestris Sond.	J. Schönenberger 366 (Z), (BOL)	AM235628	AM235396	AY151609[a]	AM235495	AM235567
Endonema lateriflora (L.f.) Gilg	J. Schönenberger 369 (Z), (BOL)	AM235629	AM235397	AY151610[a]	AM235496	AM235568
Endonema retzioides Sond.	J. Schönenberger 370 (Z), (BOL)	AJ605088[g]	AJ605101[g]	AY151611[a]	AM235497	AM235569
Glischrocolla formosa (Thunb.) R. Dahlgren	J. Schönenberger 521, photo-vouchered	AM235630	AM235398	AY151612[a]	AM235498	AM235570
Penaea acutifolia A. Juss.	J. Schönenberger 376 (Z), (BOL)	AM235631	AM235399	AY151613[a]	AM235499	AM235571
Penaea cneorum Meerb. ssp. *cneorum*	J. Schönenberger 363 (Z), (BOL)	AM235632	AM235400	AY151614[a]	AM235500	AM235572
Penaea cneorum Meerb. ssp. *gigantea* R. Dahlgren	J. Schönenberger 375 (Z), (BOL)	AM235633	AM235401	AY151615[a]	AM235501	AM235573
Penaea cneorum Meerb. cf. ssp. *lanceolata* R. Dahlgren	J. Schönenberger 320 (Z), (BOL)	AM235634	AM235402	AY151616[a]	AM235502	AM235574
Penaea cneorum Meerb. ssp. *ovata* (Eckl. & Zeyh. ex A. DC.) R. Dahlgren	J. Schönenberger 374 (Z), (BOL)	AM235635	AM235403	AY151617[a]	AM235503	AM235575
Penaea cneorum Meerb. cf. ssp. *ruscifolia* R. Dahlgren	J. Schönenberger 368 (Z), (BOL)	AM235636	AM235404	AY151618[a]	AM235504	AM235576
Penaea dahlgrenii Rourke	J. Schönenberger 388 (Z), (BOL)	AM235637	AM235405	AY151619[a]	AM235505	AM235577
Penaea mucronata L.	See GenBank associated references	AJ605090[g]	AF270756[c]	AY151620[a]	AM235506	AM235578
Saltera sarcocolla (L.) Bullock	J. Schönenberger 360 (Z), (BOL)	AJ605091[g]	AJ605103[g]	AY151621[a]	AM235507	AM235579
Sonderothamnus petraeus (Barker f.) R. Dahlgren	J. Schönenberger 362 (Z), (BOL)	AY078154[d]	AJ605104[g]	AY151622[a]	AM235508	AM235580
Sonderothamnus speciosus R. Dahlgren	J. Schönenberger 364 (Z), (BOL)	AM235638	AM235406	AY151623[a]	AM235509	AM235581
Stylapterus ericifolius (A. Juss.) R. Dahlgren	J. Schönenberger 372 (Z), (BOL)	AM235639	AM235407	AY151624[a]	AM235510	AM235582
Stylapterus ericoides A. Juss. ssp. *pallidus* R. Dahlgren	J. Schönenberger 355 (Z), (BOL)	AJ605092[g]	AJ605105[g]	AY151625[a]	AM235511	AM235583
Stylapterus fruticulosus (L. f.)	J. Schönenberger 359 (Z), (BOL)	AM235640	AM235408	AY151626[a]	AM235512	AM235584

Stylapterus micranthus R. Dahlgren	M. Johns, s.n., (Z)	AJ605093[g]	AJ605106[g]	AY151627[a]	AM235513	AM235585
Memecylaceae DC. (1827), nom. cons.						
Memecylon durum Cogn.	F. Rutschmann, fru 35, (BRUN), (Z)	AM235641		AM235444	AM235514	AM235586
Memecylon edule Roxb.	See GenBank associated references	AF215528[c]	AF215574[c]	AF215609[c]	n.a.	n.a.
Melastomataceae Juss. (1789), nom. cons.						
Bertolonia marmorata (Naudin)	F. Rutschmann, fru 78, cultivated in the Zurich Botanical Garden, (Z)	AM235642	AM235409	AM235445	AM235515	AM235587
Clidemia petiolaris (Schltdl. & Cham.)	Kew DNA database, 2534, Chase 2534, (K)	AM235643	AM235410	AM235446	AM235516	AM235588
Graffenrieda latifolia (Naudin) Triana	Fabian Michelangeli, FAM 794, (NY)	AM235644	AM235411	AM235447	AM235517	AM235589
Macrocentrum cristatum (DC.) Triana	M. F. Prévost (Fanchon) 4841, French Guyana, (CAY), (Z)	AM235645	AM235412	AM235448	AM235518	AM235590
Melastoma beccarianum Cogn.	F. Rutschmann, fru 74, (BRUN), (Z)	AM235646	AM235413	AM235449	AM235519	AM235591
Meriania macrophylla (Benth.) Triana	Fabian Michelangeli, FAM 829, (NY)	AM235647	AM235414	AM235450	AM235520	AM235592
Miconia donaeana Naudin	Fabian Michelangeli, FAM 727, (NY)	AM235648	AM235415	AM235451	AM235521	AM235593
Pternandra caerulescens Jack	J. Pereira, san 142201-142203, Sabah, Malaysia, (SAN), (Z).	AM235649	AM235416	AM235452	AM235522	AM235594
Pternandra echinata Jack (Metcalfe 1996)	See GenBank associated references	AF215520[c]	AF215559[c]	AF270744[c]	n.a.	n.a.
Rhexia virginica L.	T. Eriksson, cultivated in the Bergius Botanical Garden, Stockholm. See also GenBank associated references	U26334[b]	AF215587[c]	AF215623[c]	AM235523	AM235595
Tibouchina urvilleana (D.C.) Cogn.	See GenBank associated references	U26339[b]	AF272820[f]	AF322234[f]	AM235524	AM235596
Tococa guianensis Aublet	Fabian Michelangeli, FAM 703	AM235650	AM235417	AM235453	AM235525	AM235597
Myrtaceae s.l.						

Juss. (1789), nom. cons.						
Angophora costata (Gaertn.) Britten	Ed Biffin, cultivated in the Australian National Botanic Gardens, Canberra, (CANB)	AM235651	AM235418	AM235454	AM235526	AM235598
Callistemon citrinus (Curtis) Skeels	F. Rutschmann, fru 79, cultivated in the Zurich Botanical Garden, (Z)	AM235652	AM235419	AM235455	AM235527	AM235599
Eucalyptus lehmannii (L. Preiss ex Schauer) Benth.	F. Rutschmann, fru 80, cultivated in the Zurich Botanical Garden, (Z)	AM235653	AM235420	AM235456	AM235528	AM235600
Eugenia uniflora L.	F. Rutschmann, fru 81, cultivated in the Zurich Botanical Garden, (Z)	AM235654	AM235421	AM235457	AM235529	AM235601
Kunzea vestita Schauer	F. Rutschmann, fru 82, cultivated in the Zurich Botanical Garden, (Z)	AM235655	AM235422	AM235458	AM235530	AM235602
Leptospermum scoparium JR Forst. & G. Forst.	F. Rutschmann, fru 83, cultivated in the Zurich Botanical Garden, (Z)	AM235656	AM235423	AM235459	AM235531	AM235603
Lophostemon confertus (R. Br.) PG Wilson & Waterhouse	Ed Biffin, cultivated in the Australian National Botanic Gardens, Canberra, (CANB)	AM235657	AM235424	AM235460	AM235532	AM235604
Melaleuca alternifolia (Maid. and Bet.) Cheel	F. Rutschmann, fru 84, cultivated in the Zurich Botanical Garden, (Z)	AM235658	AM235425	AM235461	AM235533	AM235605
Metrosideros excelsa Banks. ex. Gaertn.	F. Rutschmann, fru 85, cultivated in the Zurich Botanical Garden, (Z)	AM235659	AM235426	AM235462	AM235534	AM235606
Tristaniopsis sp. indet.	F. Rutschmann, fru 70, (BRUN), (Z)	AM235660	AM235427	AM235463	AM235535	AM235607
Uromyrtus metrosideros (Bailey) AJ. Scott	Ed Biffin 9102545, cultivated in the Australian National Botanic Gardens, Canberra, (CANB)	AM235661	AM235428	AM235464	AM235536	AM235608
Heteropyxis natalensis Harv.	Peter Wilson, cultivated in the Royal Botanical Gardens, Sydney, RBG 781154.	AM235662	AM235429	AM235465	AM235537	AM235609
Psiloxylon mauritianum Thouars ex Tul.	Dennis Hansen, La Réunion, (Z)	AM235663	AM235430	AM235466	AM235538	AM235610
Vochysiaceae A. St.-Hil. (1820), nom. cons.						

Ruizterania albiflora (Warming) Marcano-Berti	Kew DNA database, 3622, RIB 1498, (INPA)	AM235664	AM235431	AM235467	AM235539	AM235611
Vochysia tucanorum Mart.	Kew DNA database, 1054, Litt 32, (NY)	AM235665	AM235432	AM235468	AM235540	AM235612
Onagraceae Juss. (1789), nom. cons.						
Circaea lutetiana L.	F. Rutschmann, fru 86, cultivated in the Zurich Botanical Garden, (Z)	AM235666	AM235433	AM235469	AM235541	AM235613
Fuchsia paniculata Lindl.	F. Rutschmann, fru 87, cultivated in the Zurich Botanical Garden, (Z)	AM235667	AM235434	AM235470	AM235542	AM235614
Fuchsia procumbens R.Cunn. ex A.Cunn.	T. Eriksson, cultivated in the Bergius Botanical Garden, Stockholm	AM235668	AM235435	AM235471	AM235543	AM235615
Gaura lindheimeri Engelm. & A. Gray	F. Rutschmann, fru 88, cultivated in the Zurich Botanical Garden, (Z)	AM235669	AM235436	AM235472	AM235544	AM235616
Ludwigia palustris	F. Rutschmann, fru 89, cultivated in the Botanischer Garten Basel, (Z)	AM235670	AM235437	AM235473	AM235545	AM235617
Oenothera macrocarpa Nutt.	F. Rutschmann, fru 90, cultivated in the Zurich Botanical Garden, (Z)	AM235671	AM235438	AM235474	AM235546	AM235618
Lythraceae Jaume St.-Hil. (1805), nom. cons.						
Cuphea hyssopifolia Kunth	F. Rutschmann, fru 91, cultivated in the Zurich Botanical Garden, (Z)	AM235672	AM235439	AM235475	AM235547	AM235619
Duabanga grandiflora (Roxb. ex DC.) Walpers	Peter Wilson, cultivated in the Royal Botanical Gardens, Sydney, RBG 811005.	AM235673	AM235440	AM235476	AM235548	AM235620

Table 2. Fossils used in this study with corresponding ages, locations, and references. The fossil ages highlighted in bold were used for calibration.

Fossils	Fossil ages	Locations	References
1. Melastomataceae leaves	**Early Eocene (53 mys)**	North Dakota	Hickey, 1977
2. Melastomeae seeds	**Miocene (26 - 23 mys)**	Russia (Siberia, Tambov region), Belarus, Poland, Germany, Belgium	Dorofeev, 1960, 1963, 1988; Collinson and Pingen, 1992; Dyjor *et al.*, 1992; Fairon-Demaret, 1994; Mai, 1995, 2000
3. Pollen of *Myrtaceidites lisamae* = *Syncolporites lisamae*	**Santonian (86 mys)**	Gabon	Herngreen, 1975; Boltenhagen, 1976; Muller, 1981
	Lower Senonian (89 – 83.5 mys)	Borneo	Muller, 1968
	Maastrichtian (71.3 – 65 mys)	Colombia	Van der Hammen, 1954
4. Fruits and seeds of *Paleomyrtinaea*	**Late Paleocene (56 mys)**	North Dakota	Crane *et al.*, 1990; Pigg *et al.*, 1993
	Early Eocene (54 mys)	British Columbia	Manchester, 1999
5. Eucalypt-like fruits	**Middle Eocene (48 mys)**	Redbank Plains Formation, Queensland, Australia	Rozefelds, 1996
	Middle Eocene (48 mys)	Lake Eyre, Nelly Creek, northern South Australia	Christophel *et al.*, 1992
6. *Fuchsia* pollen *Diporites aspis*	**Early Oligocene (33.7 – 28.5 mys)**	Australia	Daghlian *et al.*, 1985; Berry *et al.*, 1990; Berry *et al.*, 2004

Table 3. Characteristics of the 72 different calibration sets, each consisting of six calibration points: Average squared deviation *s* (see Figure 2); nodal distances summed up over all fossil calibration runs (see Figures 1 and 2); percent deviation of *s* from regression line with nodal distances (see Figures 3 and 4); mean ages of node X (see Figure 1) with standard deviations and 95% credibility intervals (see Figures 4, 5, and 6). The three sets with the highest level of corrected calibration–set consistency are shaded in light grey, whereas the three sets with the lowest corrected consistency are shaded in dark grey. Note that fossil 2 could be assigned to only one node in our phylogeny (node *2*).

Calibration set	**Calibration points**						**Average squared deviation *s***	**Nodal distances**	**Percent deviation of *s* from regression line**	**Mean age of node X [mys]**	**SD**	**95% credibility interval**	
												2.5th percentile	**97.5th percentile**
	1	***2***	***3***	***4***	***5***	***6***							
1	*a*	*2*	*a*	*a*	*a*	*a*	0.0610	214	15.870	72.80	7.92	57.25	88.81
2	*a*	*2*	*a*	*b*	*b*	*b*	0.1411	244	-17.133	78.14	7.57	63.14	92.22
3	*a*	*2*	*a*	*a*	*b*	*b*	0.1098	234	-15.960	76.66	7.62	61.30	91.12
4	*a*	*2*	*a*	*b*	*a*	*a*	0.0982	224	9.053	75.83	7.74	60.72	91.06
5	*a*	*2*	*a*	*a*	*a*	*b*	0.0769	224	-16.133	74.40	7.88	59.24	89.71
6	*a*	*2*	*a*	*b*	*b*	*a*	0.1291	234	1.419	77.27	7.67	61.27	92.36
7	*a*	*2*	*a*	*a*	*b*	*a*	0.0957	224	6.689	74.80	7.78	59.17	89.99
8	*a*	*2*	*a*	*b*	*a*	*b*	0.1120	234	-13.642	77.15	7.45	62.66	91.82
9	*a*	*2*	*b*	*a*	*a*	*a*	0.0662	216	11.062	75.55	7.88	59.63	90.29
10	*a*	*2*	*b*	*b*	*b*	*b*	0.1375	246	-25.719	79.14	7.34	64.44	93.31
11	*a*	*2*	*b*	*a*	*b*	*b*	0.1088	236	-23.954	78.15	7.49	62.97	91.98
12	*a*	*2*	*b*	*b*	*a*	*a*	0.1008	226	3.864	77.12	7.60	61.73	91.73
13	*a*	*2*	*b*	*a*	*a*	*b*	0.0792	226	-22.426	76.61	7.54	61.60	91.34
14	*a*	*2*	*b*	*b*	*b*	*a*	0.1285	236	-4.954	78.38	7.66	63.06	92.81
15	*a*	*2*	*b*	*a*	*b*	*a*	0.0977	226	0.865	76.81	7.69	61.21	91.95
16	*a*	*2*	*b*	*b*	*a*	*b*	0.1116	236	-20.846	78.19	7.43	62.86	92.43

17	*a*	*2*	*c*	*a*	*a*	*a*	0.0794	218	16.299	77.08	7.80	61.21	91.72
18	*a*	*2*	*c*	*b*	*b*	*b*	0.1417	248	-27.351	79.70	7.55	64.47	94.76
19	*a*	*2*	*c*	*a*	*b*	*b*	0.1169	238	-21.903	79.21	7.50	63.69	93.02
20	*a*	*2*	*c*	*b*	*a*	*a*	0.1102	228	5.136	78.01	7.82	61.76	92.96
21	*a*	*2*	*c*	*a*	*a*	*b*	0.0914	228	-14.361	78.50	7.57	62.56	92.77
22	*a*	*2*	*c*	*b*	*b*	*a*	0.1337	238	-6.553	78.36	7.67	62.63	92.96
23	*a*	*2*	*c*	*a*	*b*	*a*	0.1068	228	2.161	78.43	7.95	62.98	93.42
24	*a*	*2*	*c*	*b*	*a*	*b*	0.1200	238	-18.753	79.42	7.63	63.76	93.87
25	*b*	*2*	*a*	*a*	*a*	*a*	0.1414	230	20.730	72.84	8.39	56.80	89.34
26	*b*	*2*	*a*	*b*	*b*	*b*	0.2246	260	-0.661	79.30	8.13	63.07	94.74
27	*b*	*2*	*a*	*a*	*b*	*b*	0.1925	250	2.265	76.61	8.42	60.31	93.44
28	*b*	*2*	*a*	*b*	*a*	*a*	0.1794	240	16.351	75.99	8.55	59.39	92.26
29	*b*	*2*	*a*	*a*	*a*	*b*	0.1596	240	5.926	74.55	8.78	58.50	92.04
30	*b*	*2*	*a*	*b*	*b*	*a*	0.2104	250	10.601	77.48	8.59	60.92	93.88
31	*b*	*2*	*a*	*a*	*b*	*a*	0.1762	240	14.793	74.64	8.58	58.15	91.25
32	*b*	*2*	*a*	*b*	*a*	*b*	0.1955	250	3.781	77.69	8.46	60.69	93.73
33	*b*	*2*	*b*	*a*	*a*	*a*	0.1468	232	18.448	74.92	8.83	58.32	92.23
34	*b*	*2*	*b*	*b*	*b*	*b*	0.2211	262	-5.682	79.76	8.39	63.48	95.23
35	*b*	*2*	*b*	*a*	*b*	*b*	0.1916	252	-2.128	78.40	8.63	60.82	94.68
36	*b*	*2*	*b*	*b*	*a*	*a*	0.1822	242	13.427	77.53	8.54	60.81	94.36
37	*b*	*2*	*b*	*a*	*a*	*b*	0.1619	242	2.608	76.98	8.62	60.23	93.92
38	*b*	*2*	*b*	*b*	*b*	*a*	0.2099	252	6.773	78.52	8.61	61.57	95.03
39	*b*	*2*	*b*	*a*	*b*	*a*	0.1783	242	11.562	76.20	8.59	59.39	93.04
40	*b*	*2*	*b*	*b*	*a*	*b*	0.1952	252	-0.248	78.79	8.44	62.39	94.58
41	*b*	*2*	*c*	*a*	*a*	*a*	0.1599	234	20.387	76.58	8.64	59.42	93.34
42	*b*	*2*	*c*	*b*	*b*	*b*	0.2252	264	-7.131	80.43	8.41	63.90	96.26
43	*b*	*2*	*c*	*a*	*b*	*b*	0.1996	254	-1.861	78.81	8.35	61.89	94.17
44	*b*	*2*	*c*	*b*	*a*	*a*	0.1914	244	13.646	78.27	8.58	61.70	95.00
45	*b*	*2*	*c*	*a*	*a*	*b*	0.1740	244	5.026	78.43	8.62	61.70	95.19
46	*b*	*2*	*c*	*b*	*b*	*a*	0.2150	254	5.450	79.21	8.56	62.70	95.82
47	*b*	*2*	*c*	*a*	*b*	*a*	0.1873	244	11.735	77.16	8.80	60.38	94.60
48	*b*	*2*	*c*	*b*	*a*	*b*	0.2035	254	0.089	79.84	8.00	63.51	95.53

49	*c*	*2*	*a*	*a*	*a*	*a*	0.1835	240	18.205	74.49	8.62	58.29	91.65
50	*c*	*2*	*a*	*b*	*b*	*b*	0.2654	270	0.502	79.54	8.84	62.21	96.80
51	*c*	*2*	*a*	*a*	*b*	*b*	0.2335	260	3.177	76.92	8.82	59.50	94.27
52	*c*	*2*	*a*	*b*	*a*	*a*	0.2213	250	14.999	77.40	8.84	59.56	94.04
53	*c*	*2*	*a*	*a*	*a*	*b*	0.2008	250	6.307	76.19	8.79	59.28	93.81
54	*c*	*2*	*a*	*b*	*b*	*a*	0.2521	260	10.316	78.20	8.74	61.17	95.89
55	*c*	*2*	*a*	*a*	*b*	*a*	0.2181	250	13.757	76.47	8.89	58.85	93.67
56	*c*	*2*	*a*	*b*	*a*	*b*	0.2365	260	4.380	78.17	8.66	61.25	94.98
57	*c*	*2*	*b*	*a*	*a*	*a*	0.1884	242	16.280	76.51	8.58	59.66	93.18
58	*c*	*2*	*b*	*b*	*b*	*b*	0.2614	272	-3.921	80.37	8.54	63.11	96.18
59	*c*	*2*	*b*	*a*	*b*	*b*	0.2322	262	-0.655	78.37	8.96	60.34	96.05
60	*c*	*2*	*b*	*b*	*a*	*a*	0.2235	252	12.441	78.92	8.61	61.62	95.37
61	*c*	*2*	*b*	*a*	*a*	*b*	0.2026	252	3.418	77.64	8.75	60.70	95.12
62	*c*	*2*	*b*	*b*	*b*	*a*	0.2511	262	6.936	79.43	8.72	62.31	96.32
63	*c*	*2*	*b*	*a*	*b*	*a*	0.2198	252	10.947	77.35	8.80	59.40	93.97
64	*c*	*2*	*b*	*b*	*a*	*b*	0.2357	262	0.839	79.77	8.74	62.06	96.74
65	*c*	*2*	*c*	*a*	*a*	*a*	0.2013	244	17.890	77.80	8.68	59.91	94.11
66	*c*	*2*	*c*	*b*	*b*	*b*	0.2654	274	-5.245	81.46	8.69	64.39	98.24
67	*c*	*2*	*c*	*a*	*b*	*b*	0.2400	264	-0.554	79.74	8.52	63.31	96.77
68	*c*	*2*	*c*	*b*	*a*	*a*	0.2326	254	12.595	79.31	8.70	62.18	95.73
69	*c*	*2*	*c*	*a*	*a*	*b*	0.2146	254	5.256	79.12	8.90	61.32	95.80
70	*c*	*2*	*c*	*b*	*b*	*a*	0.2560	264	5.759	80.57	8.69	63.24	97.94
71	*c*	*2*	*c*	*a*	*b*	*a*	0.2285	254	11.047	78.48	8.81	61.27	95.67
72	*c*	*2*	*c*	*b*	*a*	*b*	0.2438	264	1.013	80.64	8.65	63.42	97.54

Table 4. Distribution of calibration points in calibration sets A (see Figure 3). The letter *x* in the "consensus set" indicates that, among the 12 calibration sets, the fossil was assigned to all possible calibration nodes.

Fossil	**Nodal assignments in the 12 calibration sets A**		
	a	***b***	***c***
1	all 12	0	0
2	-	-	-
3	4	4	4
4	6	6	-
5	6	6	-
6	0	all 12	-
Consensus	**1*a* 2 3*x* 4*x* 5*x* 6*b***		

Table 5. Characteristics of the six calibration sets with the highest (18, 10, 11) and the lowest (33, 41, 25) levels of corrected internal consistency.

Fossil	**Min. fossil age [mys]**	**Calibration sets with highest corrected consistency**			**Calibration sets with lowest corrected consistency**		
		set 18	**set 10**	**set 11**	**set 33**	**set 41**	**set 25**
1	**53**	*1a*	*1a*	*1a*	*1b*	*1b*	*1b*
2	**23**	*2*	*2*	*2*	*2*	*2*	*2*
3	**86**	*3c*	*3b*	*3b*	*3b*	*3c*	*3a*
4	**56**	*4b*	*4b*	*4a*	*4a*	*4a*	*4a*
5	**48**	*5b*	*5b*	*5b*	*5a*	*5a*	*5a*
6	**28.5**	*6b*	*6b*	*6b*	*6a*	*6a*	*6a*
Average squared deviation *s*		0.1417	0.1375	0.1088	0.1468	0.1599	0.1414
Percent deviation of *s* from regression line (see Figures 3 and 4)		**-27.351**	**-25.719**	**-23.954**	**18.448**	**20.387**	**20.730**
Estimated age for node X [mys] with standard deviation		79.7 ± 7.55	79.14 ± 7.34	78.15 ± 7.49	74.92 ± 8.83	76.58 ± 8.64	72.84 ± 8.39

Figures

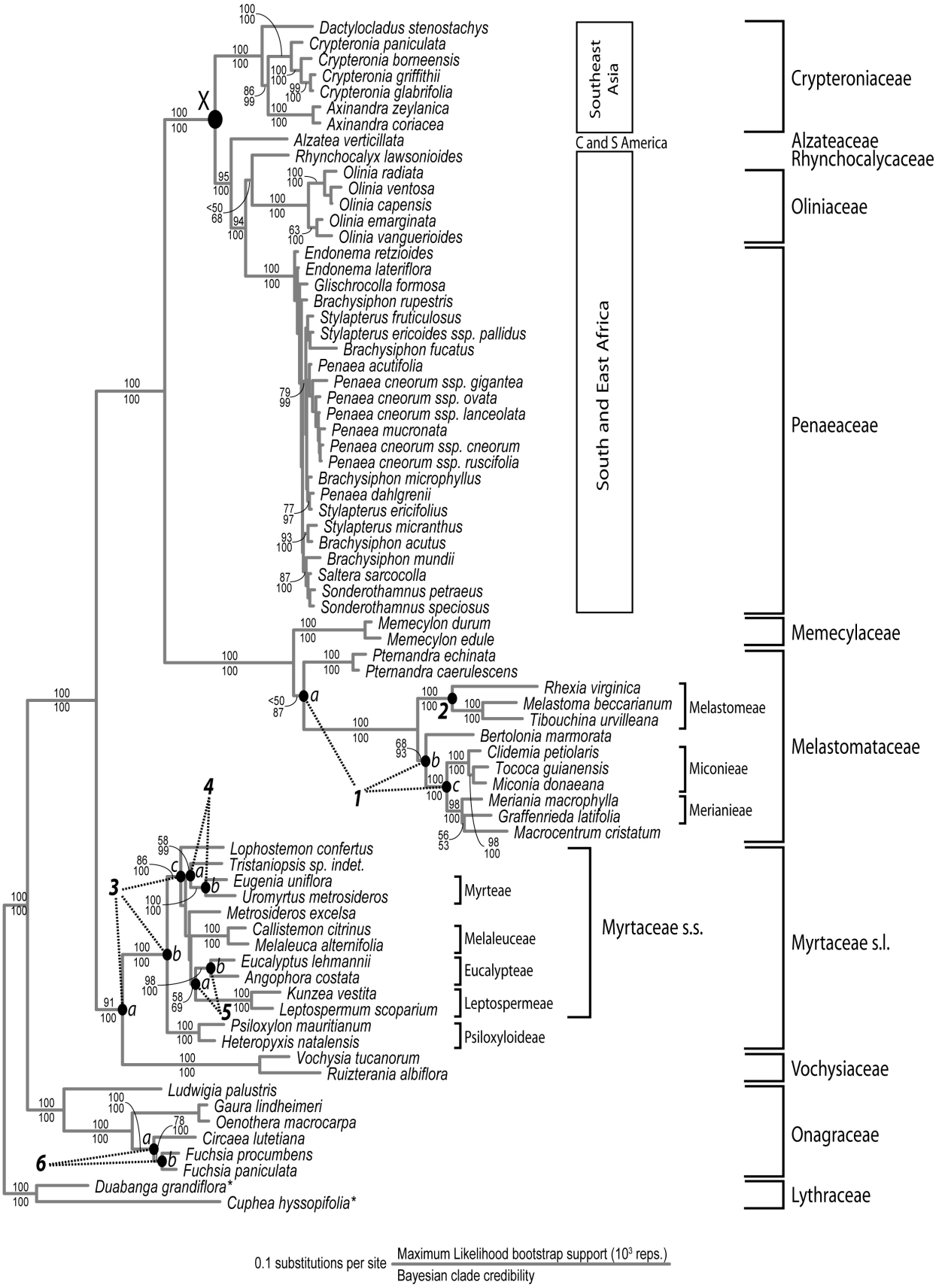


Figure 1 (see legend on next page)

Figure 1 (previous page). MrBayes majority rule consensus tree with mean branch lengths, based on the combined 5124-nucleotide data set. Maximum likelihood bootstrap support values and Bayesian clade credibility values are reported above and below the branches, respectively. General distribution ranges of the focus groups are reported to the right of the tree, as are the infra-familial ranks relevant to fossil nodal assignments. Node X represents the phylogenetic split between the Southeast Asian Crypteroniaceae stem lineage and its African/South American sister clade. Alternative nodal assignments (*a*, *b*, or *c*) for the six fossils listed in Table 2 are labelled on the tree. Outgroup taxa are indicated by an asterisk.

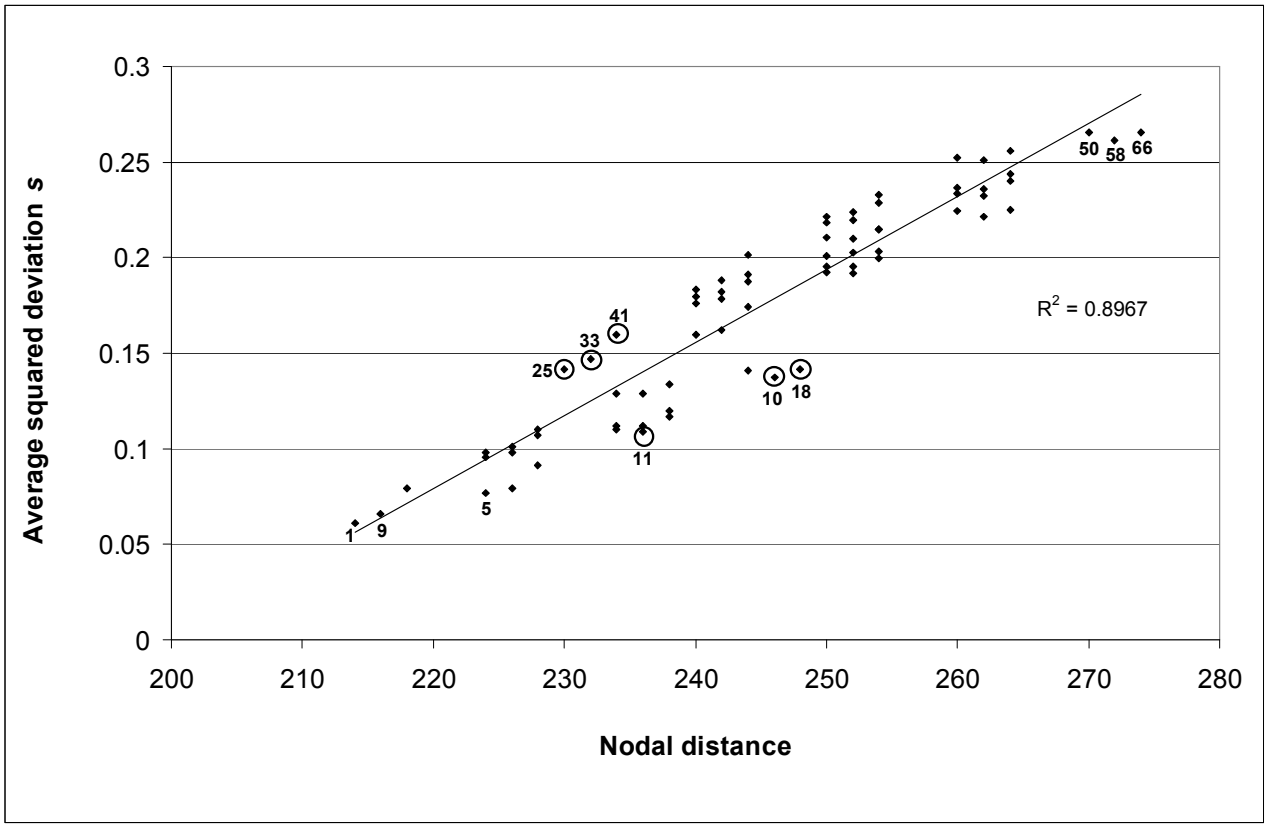


Figure 2. Correlation between the average squared deviation *s* and nodal distance among calibration points (see Table 3). R represents Pearson's correlation coefficient. Calibration sets 1, 9, 5 and 50, 66, 58 represent the sets with the lowest and highest *s* values, respectively. Calibration sets 18, 10, 11 and 33, 41, 25 (circled) show the *s* values that deviate the most below and above, respectively, the regression line.

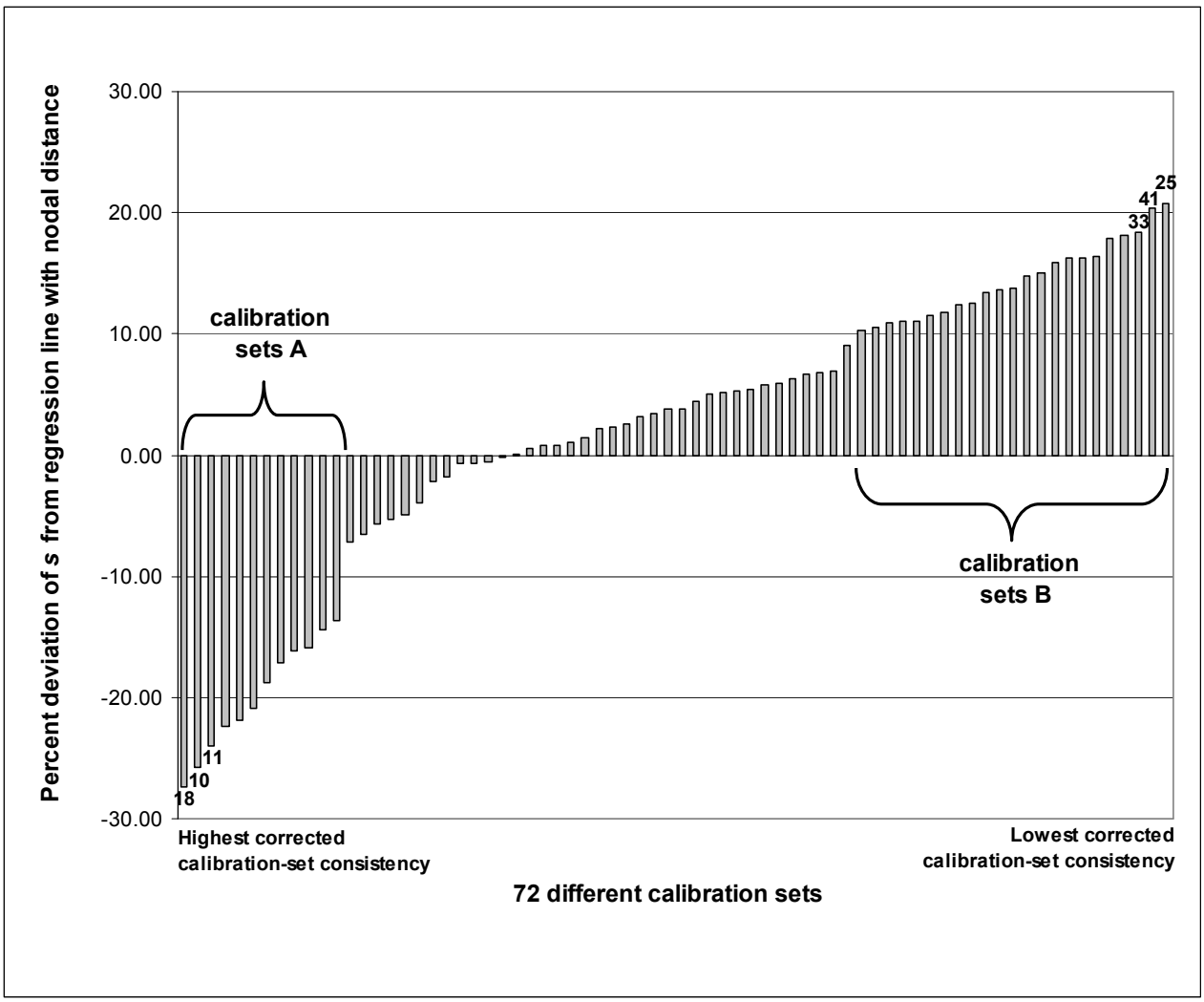


Figure 3. Histogram representing the percent deviation of *s* from the regression line of Figure 2 for all 72 calibration sets. Twelve calibration sets showed *s* values that were at least 10% lower than expected from the regression line (calibration sets A). These sets are associated with the highest level of corrected calibration-set consistency. Twenty-three calibration sets showed *s* values that were at least 10% higher than expected from the regression line (calibration sets B). These sets are associated with the lowest level of corrected calibration-set consistency.

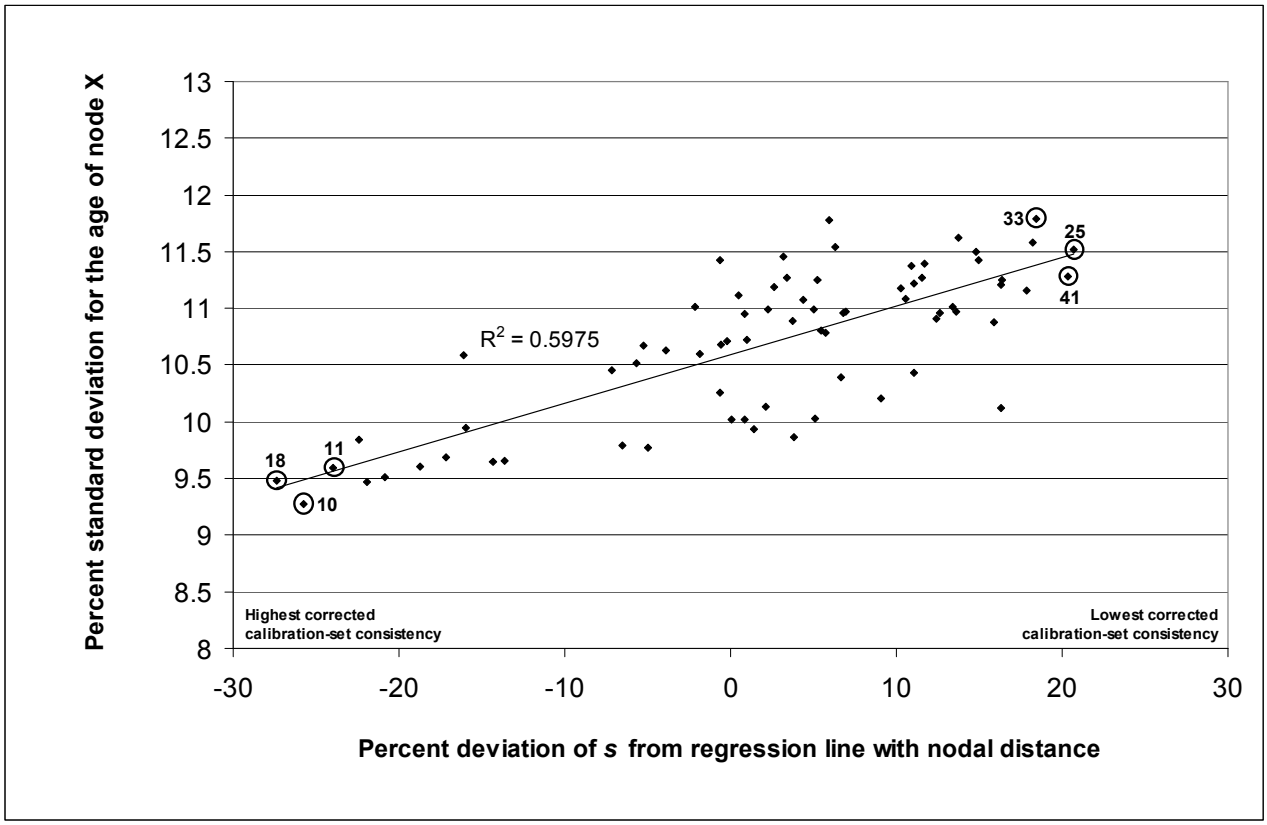


Figure 4. Correlation between corrected calibration-set consistency (expressed as percent deviation of *s* from the regression line with nodal distance; see Figures 2 and 3) and percent standard deviation for the age of node X. The three calibration sets 18, 10, and 11 produced lower standard deviations for the age of node X than the three calibration sets 33, 41, and 25.

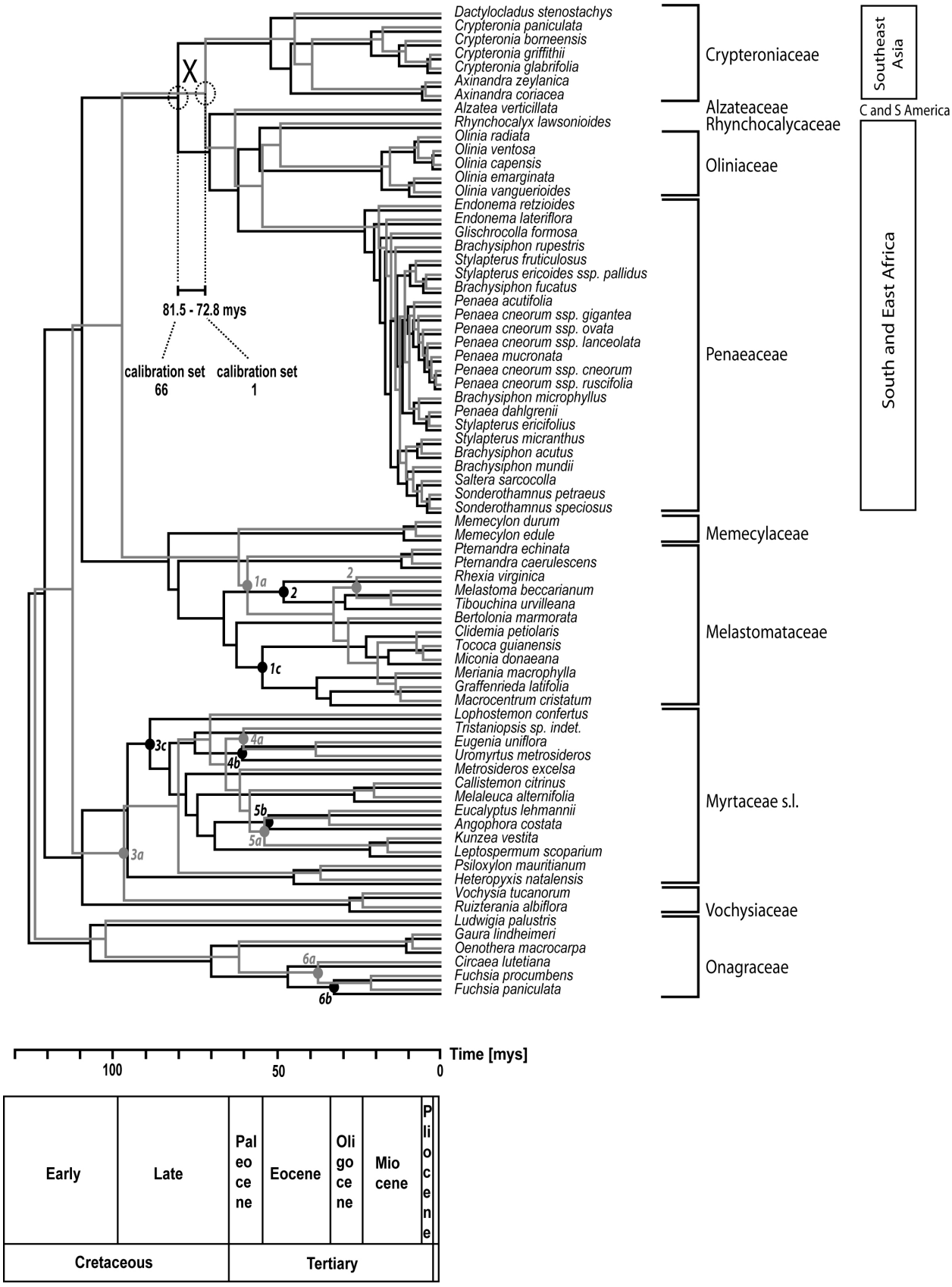


Figure 5. The two chronograms derived from the two calibration sets (66 and 1) that yielded the oldest (in black) and youngest (in grey) ages for node X, respectively. The calibration points of the two sets (set 66: *1c, 2, 3c, 4b, 5b, 6b*; set 1: *1a, 2, 3a, 4a, 5a, 6a*) are marked on the trees.

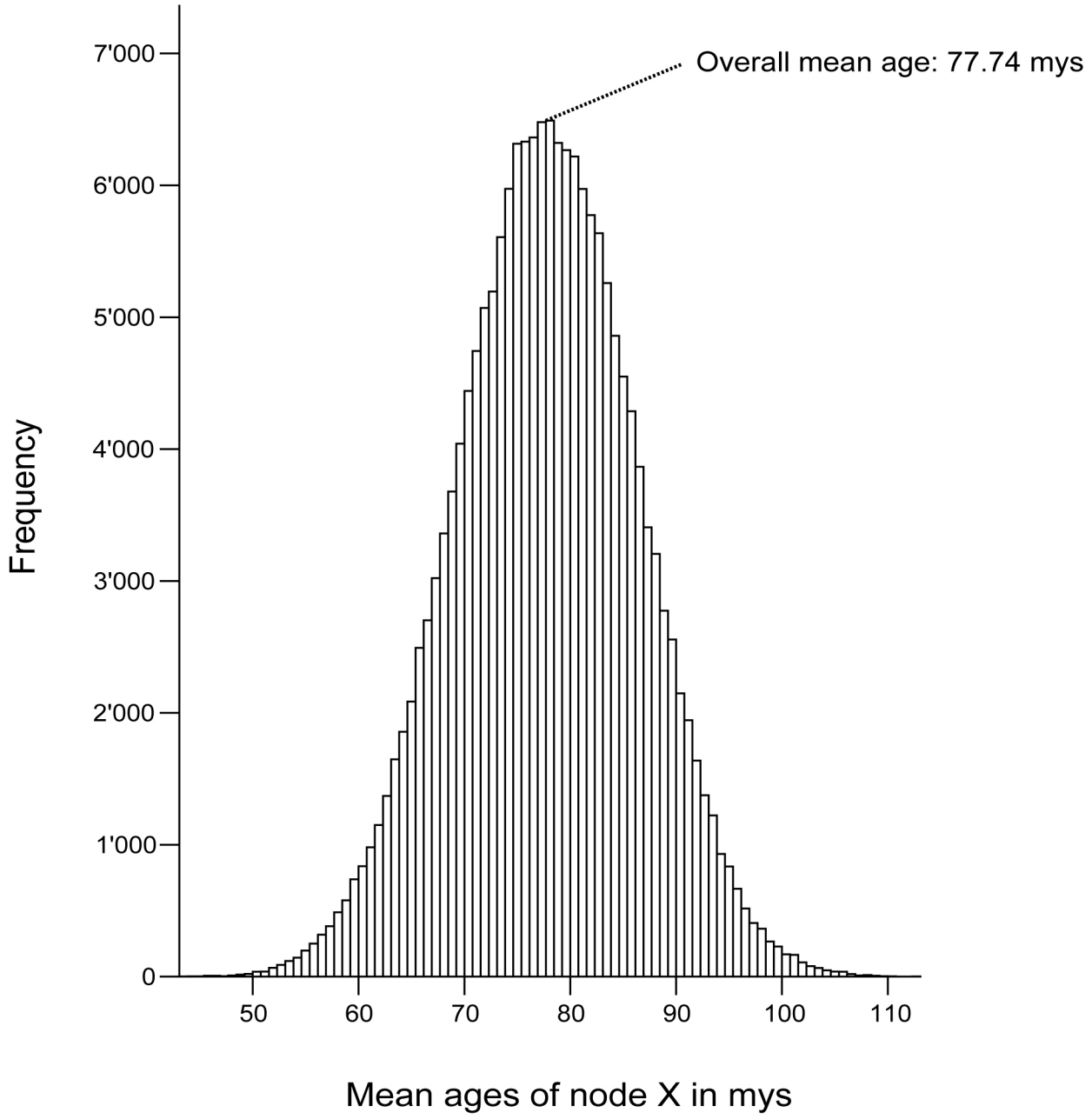


Figure 6. Distribution of mean ages for node X estimated from the 72 molecular dating analyses. Overall mean: 77.74 mys, standard deviation: 8.51 mys, 95% credibility interval (representing the 2.5th and 97.5th percentile of the posterior distribution): 60.89 – 94.21 mys.

Chapter VI.

Taxon sampling effects in molecular clock dating: An example from the African Restionaceae

H. Peter Linder[1], Christopher R. Hardy[1], and Frank Rutschmann

[1]*Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland*

Published in

Molecular Phylogenetics and Evolution 35: 569-582. 2005.

Abstract

Three commonly used molecular dating methods for correction of variable rates (non-parametric rate smoothing, penalized likelihood and Bayesian rate correction) as well as the assumption of a global molecular clock were tested for sensitivity to taxon sampling. The test dataset of 6854 basepairs for 300 terminals includes a nearly complete sample of the *Restio*-clade of the African Restionaceae (272 of the 288 species), as well as 26 outgroup species. Of this, nested subsets of 35, 51, 80, 120, 150 and the full 300 species were used. Molecular dating experiments with these datasets showed that all methods are sensitive to undersampling, but that this effect is more severe in analyses that use more extreme rate smoothing. Additionally, the undersampling effect is positively related to distance from the calibration node. The combined effect of undersampling and distance from the calibration node resulted in up to three-fold differences in the age estimation of nodes from the same dataset with the same calibration point. We suggest that the most suitable methods are Penalized Likelihood and Bayesian when a global clock assumption has been rejected, as these methods are more successful at finding optimal levels of smoothing to correct for rate heterogeneity, and are less sensitive to undersampling.

Key words: Molecular dating, NPRS, Penalized Likelihood, Bayesian dating, sampling effects, Restionaceae, Lineage through time plots.

Introduction

Dating the internal nodes of cladograms is useful for many evolutionary investigations, for example exploring plant-insect co-speciation (e.g. Percy et al., 2004), historical biogeographical analysis (e.g. Conti et al., 2002; Davis et al., 2002; Nagy et al., 2003; Vinnersten and Bremer, 2001), and relating speciation rate changes to palaeo-environmental changes (e.g. Kadereit et al., 2004; Linder, 2003). However, molecular dating is beset by a number of problems. For example, the pseudoprecision and errors that may result from the use of inadequate calibration points, and especially the use of derived calibration points which are not directly based on fossil evidence, have recently received attention (Graur and Martin, 2004; Hedges and Kumar, 2004; Lee, 1999; Shaul and Graur, 2002). Furthermore, the assumption of a global molecular clock has been shown to be invalid in many instances (Gaut, 1998). Various methods have been developed to accommodate rate variation: these include the removal of clades with deviant rates (Takezaki et al., 1995), excluding data-partitions that falsify the clock assumption (Kato et al., 2003), using several local clocks for rate-homogenous clades (i.e., the local clocks approach of Yoder and Yang, 2000), using non-parametric rate smoothing to constrain between internode rate variation (Sanderson, 1997), and searching for the optimal rates using Bayesian methods (Thorne et al., 1998) and penalized likelihood (Sanderson, 2002a). However, there seems to have been no investigation into the effects of sampling only a small proportion of the terminals (species) on the age estimates of the interior nodes. An understanding of how undersampling effects age estimates is important, as molecular phylogenetic investigations of clade ages are often based on sparse taxon samples.

Here we investigate the sensitivity of various methods of obtaining molecular age estimates to incomplete taxon sampling in the "*Restio* clade" of African Restionaceae (Poales) which, with 288 species, is the largest clade of African Restionaceae. The African Restionaceae as a whole comprise 350 species of evergreen, rush-like plants that collectively dominate much of the fynbos vegetation of the species-rich Cape Floristic Region of Southern Africa (Linder, 1991; Linder, 2003; Taylor, 1978). Specifically, we evaluate effect on node age estimates of increasing or decreasing taxon sampling, and distance from the calibration node. Our data on the *Restio* clade are particularly suited this type of investigation because (1) taxon sampling is nearly complete (ca. 95%), and (2) phylogenetic relationships are well resolved and supported by over 6,000 nucleotides of DNA sequence data.

Methods

Phylogeny estimation

Two-hundred and seventy-two species (ca. 95 %) of the 288 species of the "*Restio* clade" African Restionaceae were included in the current analysis. Additionally, both subspecies of *Restio dodii* and two accessions of the variable and widespread species *Ischyrolepis macer* were included, as they appear to represent two distinct chloroplast lineages and may be separate species. To allow the use of the basal dating node, we also included 24 species of the "*Willdenowia* clade" of African Restionaceae. As such, a total of 298 plants of African Restionaceae were sampled for this analysis. Of the 16 species of the "*Restio* clade" that were not included, three are possibly not taxonomically distinct (for detailed comment, see Linder, 2001), and the remainder could not be located in the field for the collection of extraction-quality plant material. Based on the phylogenetic studies of Briggs *et al.* (2000) and Linder *et al.* (2003), the tree was rooted to two terminals representing the ca. 150 species of Australian Restionaceae.

DNA sequences were generated from the plastid regions spanning the *trnL* intron and the *trnL-trnF* intergenic spacer (Taberlet et al., 1991), the complete gene encoding *rbcL* (Chase and Albert, 1998), the complete *atpB-rbcL* intergenic spacer (Chiang and Schaal, 2000; Cuénoud et al., 2000; Manen et al., 1994), and *matK* plus the flanking *trnK* intron (Hilu and Liang, 1997). Total DNA was isolated from silica gel-dried culms using the DNeasy® Plant Mini Kit (Qiagen, Inc.; Valencia, California, USA). Sequences were generated using standard methods for PCR amplification and automated sequencing.

Raw sequence data files were analysed with the ABI Prism™ 377 Software Collection 2.1. Contigs were constructed in Sequencher™ and alignments were performed using the default alignment parameters in Clustal X (Thompson et al., 1997), followed by adjustment by eye. These sequences were assembled into a single matrix in WinClada (Nixon, 2002). The aligned matrix consisted of 6854 aligned bases, of which 1512 are parsimony informative. Additionally, indels were coded at the end of the matrix using Simple Indel Coding (Simmons and Ochoterena, 2000) as implemented in the program GapCoder (Young and Healy, 2001). The total matrix consists of 1782 parsimony-informative characters. All characters were weighted equally and treated as nonadditive during tree searches. This data matrix has been deposited at www.treebase.org.

Parsimony searches were conducted using the parsimony ratchet (Nixon, 1999) as implemented from WinClada, running NONA vers. 1.6 (Goloboff, 1993) as a daughter process. Ten ratchet searches were conducted, each initiated with the generation of a Wagner tree, using a random taxon entry sequence, followed by TBR branch swapping with one tree retained and used

as the starting point for 500 ratchet cycles. In the weighted/constrained half of each ratchet cycle, a randomly selected set of 10% of the characters were resampled, and a randomly selected set of 10% of the resolved clades were constrained. This analysis resulted in 885 equally most parsimonious cladograms (L = 5415, CI = 0.44, RI = 0.84; informative characters only). These were then pooled and swapped to obtain a total of 10,615 cladograms of length 5415. One of these cladograms was arbitrarily chosen for the subsequent investigation into the impact of taxon sampling on the estimation of absolute dates and divergence times.

Construction of smaller subset matrices and cladograms

Using our 300 taxon matrix (not including indels) and tree as fixed starting points, six smaller matrices and trees were constructed by deleting terminals in Mesquite 1.02 (Maddison and Maddison, 2003). These smaller datasets have 150, 120, 100, 80, 51, and 35 species/terminals respectively. The list of species and sequences in each smaller set is a precise subset of the next larger set, and each employed the same relative alignment and tree topology as those obtained from the 300 taxon analysis. The only differences lie in the numbers of terminals and by the exclusion of extraneous gaps from the larger matrices that are no longer necessary in the smaller matrices. As such, each successively smaller matrix consists of 6623, 6547, 6480, 6399, 6248, and 6135 aligned bases. For the smallest (35 species) sampling, at least two representatives of the basal lineages for each of the 32 clades depicted in Figure 1 were chosen. Successively larger data matrices and cladograms simply added descendant species and, therefore, more distal nodes to these 32 nodes of interest (the "test nodes"). Thus the proportion of the descendent species sampled differs enormously among the test nodes, as does the rate at which the sampling density increases (Table 1). This particular strategy was chosen because it results in a set of comparable test nodes for each sampling set (Simmons et al., 2004). Sampling basal lineages is also the method used by phylogeneticists to estimate the age of particular clades with incomplete sampling. Only test nodes 1-30 were used in the analysis. Node 32 is the constrained basal node, and node 31 is at the base of the *Willdenowia* clade, and as such is not part of the study group.

NPRS, PL and clock analyses

As preparation for the clock assumption (CL), non-parametric rate smoothing (NPRS, Sanderson, 1997), and penalized likelihood (PL, Sanderson, 2002a) approaches to dating, molecular branch lengths were estimated for each of the seven nested matrices and cladograms. We used the implementation of Modeltest (Posada and Crandall, 1998) in Hy-Phy ver. 0.99 beta for Windows® (Kosakovsky Pond et al., 2004) to select a statistically adequate model from a set

of 56 possible models of sequence evolution. Using the selected models (Table 2), likelihood ratio tests (Felsenstein, 1981) were performed in Hy-Phy to test for a significant departure from the hypothesis of a global molecular clock. In each case, the clock was rejected (Table 2). Branch lengths were estimated in PAUP* 4.0 (Swofford, 2002) using the appropriate model without a clock assumption.

These branches were made ultrametric using NPRS, as implemented in TreeEdit for Macintosh (Rambaut and Charleston, 2004), and penalized likelihood with r8s (ver. 1.6 for Linux, Sanderson, 2002b). For the latter, an optimal rate-smoothing parameter value was selected with the prerequisite cross-validation procedure (Sanderson, 2002a) for all except the largest two matrices (i.e., 150 and 300 terminals). Attempts to find an optimal smoothing parameter for the 150 and 300 taxon sets failed, possibly due to computational limitations with these large datasets, or possibly because of the presence of zero-length terminal branches. For the 120 taxon and fewer datasets, smoothing parameter values ranging from $l0^{-3.5 \text{ to } 7.5}$ (in increments of $10^{0.5}$) were tested and the resulting values reported in Table 2. In order to compare the four methods, the optimal regression of the taxon sampling and average node age values for the 35 to 120 sample PL method were used to predict the values for the 150 and 300 taxon samples. These predicted values were not included in any statistical testing.

Despite rejecting the clock, branch lengths were made ultrametric also under the assumption of a molecular clock, for comparative purposes, using the appropriate model in PAUP*.

We calibrated the trees against node 32 (connecting the *Restio* and *Willdenowia* clades). We used a secondary date (49.8 Ma, with a range of 42.7- 55.9 Ma), obtained from the analysis of Linder *et al.* (2003). In this analysis the adjacent node (connecting the African and Australian Restionaceae) was dated from an African pollen deposit from the earliest Tertiary (Linder et al., 2003; Scholtz, 1985).

Bayesian dating

The Bayes dating method (Thorne and Kishino, 2002; Thorne et al., 1998) uses a probabilistic model to describe the change in evolutionary rate over time and uses the Markov chain Monte Carlo (MCMC) procedure to derive the posterior distribution of rates and time. It allows multiple calibration windows and provides direct credibility intervals for estimated divergence times and substitution rates. The procedure we followed is divided into three different steps and programs, and is described in more detail in a step-by-step manual available at http://www.plant.ch/software.html. It was performed on a 3 Ghz Pentium IV machine running Windows XP. In a first step, we estimated the model parameters for the the F84 + G model

(Felsenstein, 1993; Kishino and Hasegawa, 1989), the most complex model of nucleotide substitution implemented in the software below so far. By using the program *Baseml*, which is part of the PAML package (Yang, 1997), we estimated base frequencies, transition / transversion rate kappa, and the alpha shape parameter (describing the rate heterogeneity among sites under a discrete gamma model; 5 categories of rates). Then, by using these parameters, we estimated the maximum likelihood of the branch lengths of the rooted evolutionary tree together with a variance-covariance matrix of the branch length estimates by using the program *Estbranches* (Thorne et al., 1998). The maximum likelihood scores obtained in *Baseml* and *Estbranches* were then compared to check if both approaches were able to optimize the likelihood. The third program we used, *Multidivtime* (Kishino et al., 2001; Thorne and Kishino, 2002), approximates the posterior distributions of substitution rates and divergence times by using a multivariate normal distribution of estimated branch lengths (provided here by *Estbranches*) and running a Markov chain Monte Carlo procedure. Two constraints for the age of node 32 were set: a lower constraint of 42.7 Ma, and a higher one of 55.9 Ma, representing the extreme values obtained for this node by Linder *et al.* (2003). The other settings for the prior distributions were: 50 for both *rttm* (mean of the prior distribution for the time separating the ingroup root from the present) and *rttmsd* (the prior's standard deviation), 0.004 for both *rtrate* (mean of the prior distribution for the rate of molecular evolution at the ingroup root node, calculated by taking the mean distance between the ingroup root and the ingroup tips obtained from estbranches) and *rtratesd* (the prior's standard distribution). *Brownmean* (the mean of the prior distibution for the Brownian motion parameter „nu", which determines the permitted rate change between ancestral and descendant nodes) was initially left at the default value of 0.4. Later, we changed that value to 0.02 and repeated the analysis, following the manual's recommendation that *rttm* multiplied with „nu" should be about 1. As this did not affect the divergence time estimates significantly, we report here only the results from the first analysis. *Brownsd*, the prior's standard deviation was chosen to be 0.4. For the parameter *bigtime*, a number that should be set higher than the time units between the tips and the root in the user's wildest imagination, we've chosen a value of 100.

We ran the Markov chain for at least 10^4 cycles and collected one sample every 100 cycles, after an unsampled burnin of 10^4 cycles. We performed each analysis at least twice by using different initial conditions to assure convergence of the Markov chain, although it is not possible to say with certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution (Cowles and Carlin, 1996).

Statistical testing and lineage through time plots

The hypothesis that the test node ages obtained are related to the number of taxa sampled was tested using the Wilcoxon Paired Sample test, which compares the number of instances in which the larger sample finds an older date compared to the number of times it finds a younger date. The hypotheses that changes in age estimation are related to the distance from the calibration point, and to the degree of sampling of the subclade subtended by each test node, were evaluated using linear regressions. This allowed us to statistically test both the extent to which the variation in the data was accounted for by the regression line, and also whether the slope of the regression line deviates significantly from horizontal. All statistical tests were conducted using SPSS. For each analysis a lineage through time (LTT) plot (Nee et al., 1992) was constructed. The rate constancy of the radiation in Restionaceae was tested using the Constant Rates test of Pybus and Harvey (2000), as implemented in Gammastatistic v1.0 (Griebeler, 2004).

Results

Undersampling

All four methods find more or less the same ages for the 30 test nodes when only 35 taxa are sampled. However, when more taxa are sampled, the age estimates of the test nodes diverge rapidly (Figure 2, Table 3). The proportion of species sampled (thus the proportion of nodes distal to the test nodes) clearly has a major impact on the ages estimated for the test nodes (Figure 3). For all four methods the mean estimated ages of the nodes are significantly less with a sparser taxon sampling than when all taxa are included in the final calculation. Thus, not including all taxa in the sample results in a "younger" estimation of the test node ages. Furthermore, for all four methods the degree of age underestimation increases logarithmically with the proportion of undersampling (Figure 3).

However, taxon undersampling has very different effects in the four methods. The CL analysis and PL are only slightly affected: for the 35 taxon samples the average ages using CL are 91%, and using PL 88%, of those obtained with the 300 taxon sample. The regression line explains only 73% of the variation in these data for CL, and 72% in the case of PL. Whilst the more severe undersampling in both CL and PL resulted in significant age change, in both there is no significant change in the age estimates between the 150 and 300 taxon samples for CL, suggesting that at 50% sampling an asymptote had been reached, at least for CL (the values for the 150 and 300 taxon samples were inferred for PL, and so the asymptote cannot be calculated).

For Bayesian and NPRS analyses the effects are more dramatic, and no age asymptote is reached. Thus any change in sampling resulted in significantly different age estimations. For NPRS the 35 taxon sample the average age estimate is only 56% of the estimate with the 300 taxon dataset, and for the fitted regression line $r^2 = 0.9804$. For Bayesian analysis the equivalent values are 72% and $r^2 = 0.9819$.

The regular decrease in the age estimations with decreasing sample size is reflected in the very good fit of the data to the logarithmic regression, and indicates that these patterns are not random.

Distance from the calibration point

For NPRS and the Bayesian analysis there is a significant (at $p < 0.01$), positive, linear relationship between the degree of underestimation of the age of a node and its distance (time) from the calibration point, for the 35 taxon sample (Figure 4). The slope of the regression is somewhat steeper for NPRS than for the Bayesian analysis, and also explains more of the

variation ($r^2 = 0.9087$ compared to $r^2 = 0.7534$). In both these cases the slope deviates significantly from the horizontal. Thus the more distant a node is from the calibration point, the more sensitive its age estimation is to the effects of undersampling. And conversely, the closer it is to its calibration point, the less sensitive it is to taxon undersampling. For the NPRS 35-taxon analysis, the most distant test node from the calibration point is dated to only 37.7% of the age indicated by the 300 taxon sample (7 instead of 18.7 Ma for node 5). These effects are much less severe in the Bayesian analysis, where this node is dated to 6.6 instead of 10.4 Ma (an underestimation of 63%). Conversely, the most proximal node in the NPRS 35 taxon analysis is estimated to be 79% of the value of the 300 taxon sample (32.96 instead of 41.66 Ma).

The CL analysis and PL are less sensitive to this distance effect, and in their cases the regression explains very little of the variation ($r^2 = 0.1469$ and $r^2 = 0.0344$ respectively). Interestingly, both these analyses show a much wider scatter, consistent with the assumption of a more clock-like molecular variation. For both these analyses neither the variation explained, nor the deviation of the regression slope from horizontal, is significant.

Impact of undersampling of individual clades

The sensitivity of the various methods to variation in the sampling density of the clades subtended by each evaluated node cannot be rigorously tested from our data, since most of these clades were rather poorly sampled in the 35 taxon sample. For most clades less than 30 percent of the species were included, one clade includes 50% of the species, one 63%, and one all species. Nonetheless, it appears as if there is no relationship between the sampling density of the individual clades, and the age estimation of their subtending nodes (Figure 5) for any of the four methods used. Neither the variance in the data, nor the deviation of slope of the regression line from horizontal, is significant, suggesting that the subclade sampling has no impact on the results. Thus only the average sampling density on the whole tree under investigation has an impact on the results, not the sampling density of the individual subclades.

Age estimates by the four methods

The four methods result in rather different age estimates (Table 3, Figure 6) for the 300 taxon data set. As expected, the extremes are formed by CL and NPRS. The NPRS analysis returns results that are between 10 and 15 million years older (thus up to double) the age estimates of the CL analysis. This suggests that all the nodes have been made older, and this could only be achieved by interpreting the basal branch of the whole tree, between the calibration point and the first speciation events, as being much shorter than the unsmoothed data indicates. Bayesian analysis and PL return intermediate ages for the nodes. However, for the nodes further

from the calibration point, Bayesian ages approach those of PL and CL, while nodes closer to the calibration point are more intermediate. The most extreme disparity is found in the middle section of the tree, for example node 10, which CL dates as 9.28, Bayesian as 14.04 and NPRS as 22.62 Ma, thus more than two-fold differences.

The LTT plots for the four methods are remarkably different (Figure 7). In all analyses the nodes are shifted towards the base of the tree, indicating that with time the speciation rate slowed down. For NPRS the shift is highly significant ($p < 0.01$), for CL and PL weakly significant ($p < 0.05$), and for Bayesian it is not significant.

Discussion

The differences in the node ages reported by the different methods, and for different sampling intensities, are remarkably large. On average, the highest age estimate for each node is 2.09 times larger than the smallest estimate, this factor ranges from a minimum of 1.41 to a maximum of 2.94. This indicates a potentially substantial source of error for dating studies. This error is determined by both the overall sampling density and the distance from the calibration node.

Such large differences have been reported before. For example, Klak *et al.* (2004) reported a two fold difference in the age estimation of the start of the radiation time of the African Rushioideae for two different genes; however, one gene was sampled for twice as many species as the other gene. Our results point to the possibility that this discrepancy may not due to differences in the molecular evolution of the two genes, but to differences in taxon sampling.

Undersampling

Undersampling has a severe impact on the results in rate-smoothed analyses, and with increasing undersampling the impact rapidly becomes more extreme. With NPRS, a taxon sampling of 10% can result in age estimates that are half of the correct value (assuming that sampling all species gives the correct value). The fitted curve to the undersampling effect indicates that it is logarithmic, which means that increasing the undersampling might increasingly rapidly exacerbate the age underestimation. Many recent studies included less than 10% of their species, suggesting that they could be prone to the undersampling error.

Although sampling significantly effects analyses that have limited (PL) no (CL) rate smoothing, this effect in our studies was less than 15%. Experiments with PL, involving changing the smoothing parameter, showed that decreasing the smoothing function λ comes with the cost of increased sensitivity to sampling effects (data not shown). It therefore does not automatically follow that the use of PL would eliminate sampling effects. In addition, we do not know what happens when the sampling is less than 10%, and would caution against using these results to suggest that in all circumstances undersampling can be accommodated by using CL or PL.

It is most likely that the effect is not due to simple undersampling. In our experiments we added only nodes that were distal to the test nodes (relative to the calibration node), and it is possible that if nodes were added both proximally and distally to the test nodes, there might be no undersampling effect. Adding only proximal nodes might result in the test nodes being shifted down the tree (further into the past). This shifting of the nodes is manifested as changing age estimates. However, it is difficult to avoid biased sampling. Unbiased sampling is only possible if

we know the relative time positions of all the nodes, and can select them to keep the proportions of proximal and distal nodes equal. Unfortunately we do not know their relative positions without first including all of them in a dating analysis. Random sampling does not help, since randomly selected species bias towards retrieving the deeper nodes (Pybus and Harvey, 2000). Furthermore, most investigators are interested in establishing the age of a preset number of nodes, deep inside the tree (e.g. the starting age of the radiation of Rushioideae (Klak et al., 2004), Angiosperms (Sanderson and Doyle, 2001; Wikström et al., 2001), *Phylica* (Richardson et al., 2001)) and these nodes can only be retrieved if the two basal-most descendents of the node are included in the analysis. This forces an unbalanced sampling.

Furthermore, the vagaries of extinction and speciation are not likely to have left a temporally evenly spaced set of surviving taxa. Clusters of short branch lengths in a phylogeny have been reported in very divergent groups, such as commelinid monocots (Duvall et al., 1993) and sponge-dwelling snapping shrimps (Morrison et al., 2004). Consequently we cannot establish what a full species sample constitutes, and so the extent of undersampling cannot be determined. Thus only methods that would minimize the undersampling effect can be considered to be robust. Since all methods we tried showed some undersampling effect, it might be useful to search for an asymptote, as was shown by the CL and to a lesser extent the PL methods.

Distance from the calibration node

There is a remarkably linear relationship between the degree of under estimation of the test node ages, and the distances from the calibration point for NPRS and the Bayesian analysis. Although a weak trend is visible, there is no significant linear relationship for the CL and PL analyses.

This argues, at least in analyses using NPRS or Bayesian analysis, that the calibration nodes should be situated within the study group, as has been suggested by Shaul and Graur (2002). Arguments for multiple calibration points are usually to protect against errors in single calibration points (Lee, 1999), and a second strong argument is that dated nodes are so placed in the proximity of fixed nodes, thus reducing the error that might accumulate over longer time spans.

Impact of undersampling of individual nodes

Somewhat surprisingly, we show that the age estimate for a test node is not affected by how complete the sampling is for the clade subtended by the node. Instead, the average level of sampling for the whole data set is of importance. Thus the sampling effect cannot be avoided by sampling one clade exhaustively, and leaving a number of place-holders for the rest of the study

group. Clearly the effects of sampling density are spread more or less evenly across the ingroup. This is advantagous in that groups that are species poor, possibly due to extinction, are not intrinsically impossible to date, but it does argue against strategies of sampling a group of interest in detail, while the related groups are undersampled.

Age and rate of diversification of Restionaceae

As the sampling becomes better, the age estimates for the African Restionaceae diverge. Thus more and better data do not result in a convergence to a single possibly correct answer. This indicates that not only taxon sampling, but also the choice of algorithm, is important.

The clock assumption (CL) is fairly robust to undersampling and to distance from the calibration node, and should therefore be used whenever possible. However, if the clock assumption is rejected, as it was for our Restionaceae data, CL cannot be used. On the LTT plot, CL results in a "wobbly" line (Figure 7), which could either be interpreted as a variable net diversification rate, or as violations of the clock. Since the latter has been demonstrated, we can ignore these results.

NPRS returned remarkably divergent results from the other methods with respect to the estimated age of the nodes. Its great sensitivity to both sampling effects and the distance from the calibration node suggest that the danger of over-smoothed results is real. More remarkable is the effect of NPRS on the LTT plot. Not only does the radiation start earlier than predicted by the other methods, but the initial phases of the radiation are interpreted to be much more rapid than by the other methods, so that a constant rates test shows a significant change in the net diversification rate. These results may be due to "over-smoothing" (Sanderson, 2002a).

Bayesian and PL were the most resilient to undersampling with our Restionaceae data. Both methods are computationally very intensive, and our 300 taxon Bayesian analysis required more than four weeks of computation time. Possibly PL is the best, since it is not sensitive to the distance from calibration, but we were not able to complete the cross-validation analyses to obtain the optimal smoothing values for the PL analyses of 150 and 300 taxa. The only difference between the two methods is the estimation of the start date of the radiation. Thus, when a global clock assumption is rejected, we recommend that either PL or Bayesian analyses should be used.

Beyond the Restionaceae

It is difficult to generalize from our results, since they are based on the results of a single study. Generality can be achieved by using simulated data sets, but then we don’t know how well simulations will mimic the real situation. We have not attempted to evaluate our results with simulated data sets, largely because of the enormous computing effort that would be needed to

analyse a sufficiently large set of replicates. However, our results clearly demonstrate that caution is required when using rate-smoothing methods, and that an understanding of the potential effects of sampling and calibration position on age estimates is a pre-requisite to any study.

References

Briggs, B. G., A. D. Marchant, S. Gilmore, and C. L. Porter. 2000. A molecular phylogeny of Restionaceae and allies. Pp. 661-671 *in* K. L. Wilson and D. A. Morrison, eds. Systematics and evolution of monocots - Vol. 1 of Proceedings of the Second International Conference on the Comparative Biology of the Monocots, Sydney, September 1998. CSIRO, Melbourne.

Chase, M. W., and V. A. Albert. 1998. A perspective on the contribution of plastid rbcL DNA sequences to angiosperm phylogenetics. Pp. 488-507 *in* D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. Molecular systematics of plants II: DNA sequencing. Kluwer Academic Publishers, Boston.

Chiang, T., and B. A. Schaal. 2000. Molecular evolution of the atpB-rbcL noncoding spacer of chloroplast DNA in the moss family Hylocomiaceae. *Bot. Bull. Acad. Sin.* **41**:85-92.

Conti, E., T. Eriksson, J. Schonenberger, K. J. Sytsma, and D. A. Baum. 2002. Early tertiary out-of-India dispersal of Crypteroniaceae: Evidence from phylogeny and molecular dating. *Evolution* **56**:1931-1942.

Cowles, M. K., and B. P. Carlin. 1996. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* **91**:883-904.

Cuénoud, P., M. A. D. Martinez, P. A. Loizeau, R. Spichiger, S. Andrews, and J. F. Manen. 2000. Molecular phylogeny and biogeography of the genus *Ilex* L. (Aquifoliaceae). *Annals of Botany* **85**:111-122.

Davis, C. C., C. D. Bell, S. Mathews, and M. J. Donoghue. 2002. Laurasian migration explains Gondwanan disjunctions: Evidence from Malpighiaceae. *Proceedings of the National Academy of Sciences of the United States of America* **99**:6833-6837.

Duvall, M. R., M. T. Clegg, M. W. Chase, W. D. Clark, W. J. Kress, H. G. Hills, L. E. Equiarte, J. F. Smith, B. S. Gaut, E. A. Zimmer, and G. H. Learn. 1993. Phylogenetic hypotheses for the monocotyledons constructed from *rbc*L sequence data. *Annals of the Missouri Botanical Garden* **80**:607-619.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **35**:292-303.

Felsenstein, J. 1993. Phylogeny Inference Package (PHYLIP). University of Washington, Seattle.

Gaut, B. S. 1998. Molecular clocks and nucleotide substitution rates in higher plants. *Evolutionary Biology* **30**:93-120.

Goloboff, P. A. 1993. NONA. Published privately, New York.

Graur, D., and W. Martin. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends in Genetics* **20**:80-86.

Griebeler, E. M. 2004. Gammastatistic V 1.0. Institut für Zoologie, Abteilung Ökologie, Johannes Gutenberg-Universität, Germany, Mainz.

Hedges, S. B., and S. Kumar. 2004. Precision of molecular time estimates. *Trends in Genetics* **20**:242-247.

Hilu, K. W., and H. P. Liang. 1997. The matK gene: Sequence variation and application in plant systematics. *American Journal of Botany* **84**:830-839.

Kadereit, J. W., E. M. Griebeler, and H. P. Comes. 2004. Quaternary diversification in European alpine plants: pattern and process. *Philosophical Transactions of the Royal Society London. B* **359**:265-274.

Kato, Y., K. Aioi, Y. Omori, N. Takahata, and Y. Satta. 2003. Phylogenetic analyses of *Zostera* species based on rbcL and matK nucleotide sequences: Implications for the origin and diversification of seagrasses in Japanese waters. *Genes & Genetic Systems* **78**:329-342.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution* **29**:170-179.

Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* **18**:352-361.

Klak, C., G. Reeves, and T. A. Hedderson. 2004. Unmatched tempo of evolution in Southern African semi-desert ice plants. *Nature* **427**:63-65.

Kosakovsky Pond, S. L., S. V. Muse, and S. D. W. Frost. 2004. Hyphy Package: Hypothesis Testing using Phylogenies

Lee, M. S. Y. 1999. Molecular clock calibrations and metazoan divergence dates. *Journal of Molecular Evolution* **49**:385-391.

Linder, H. P. 1991. A review of the southern African Restionaceae. *Contributions from the Bolus Herbarium* **13**:209-264.

Linder, H. P. 2001. The African Restionaceae. Contributions from the Bolus Herbarium, Cape Town.

Linder, H. P. 2003. The radiation of the Cape flora, southern Africa. *Biological Reviews* **78**:597-638.

Linder, H. P., P. Eldenäs, and B. G. Briggs. 2003. Contrasting patterns of radiation in African and Australian Restionaceae. *Evolution* **57**:2688-2702.

Maddison, W. P., and D. Maddison. 2003. Mesquite

Manen, J., A. Natali, and F. Ehrendorfer. 1994. Phylogeny of Rubiaceae-Rubieae inferred from the sequence of a cpDNA intergene region. *Plant Systematics and Evolution* **190**:195-211.

Morrison, C. L., R. Rios, and J. E. Duffy. 2004. Phylogenetic evidence for an ancient rapid radiation of Caribbean sponge-dwelling snapping shrimps (Synalpheus). *Molecular Phylogenetics and Evolution* **30**:563-581.

Nagy, Z. T., U. Joger, M. Wink, F. Glaw, and M. Vences. 2003. Multiple colonization of Madagascar and Socotra by colubrid snakes: evidence from nuclear and mitochondrial gene phylogenies. *Proceedings of the Royal Society of London Series B-Biological Sciences* **270**:2613-2621.

Nee, S., A. O. Mooers, and M. Harvey. 1992. Tempo and mode of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Science of the United States of America* **89**:8322-8326.

Nixon, K. C. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* **15**:407-414.

Nixon, K. C. 2002. WinClada. Published by the author, Ithaca, NY.

Percy, D. M., R. D. M. Page, and Q. C. B. Cronk. 2004. Plant-insect interactions: double-dating associated insect and plant lineages reveals asynchronous radiations. *Systematic Biology* **53**:120-127.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**:817-818.

Pybus, O. G., and P. H. Harvey. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London Series B-Biological Sciences* **267**:2267-2272.

Rambaut, A., and M. Charleston. 2004. TreeEdit: An application for organising, viewing and manipulating sets of phylogenetic trees. Published privately, Oxford.

Richardson, J. E., F. M. Weitz, M. F. Fay, Q. C. B. Cronk, H. P. Linder, G. Reeves, and M. W. Chase. 2001. Rapid and recent origin of species richness in the Cape flora of South Africa. *Nature* **412**:181-183.

Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* **14**:1218-1231.

Sanderson, M. J. 2002a. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution* **19**:101-109.

Sanderson, M. J. 2002b. r8s. Published privately, Davis.

Sanderson, M. J., and J. A. Doyle. 2001. Sources of error and confidence intervals in estimating the age of angiosperms from *rbc*L and 18S rDNA data. *American Journal of Botany* **88**:1499-1516.

Scholtz, A. 1985. The palynology of the upper lacustrine sediments of the Arnot Pipe, Banke, Namaqualand. *Annals of the South African Museum* **95**:1-109.

Shaul, S., and D. Graur. 2002. Playing chicken (*Gallus gallus*): methodological inconsistencies of molecular divergence date estimates due to secondary calibration points. *Gene* **300**:59-61.

Simmons, M. P., and H. Ochoterena. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology* **49**:369-381.

Simmons, M. P., K. M. Pickett, and M. Miya. 2004. How meaningful are Bayesian support values? *Molecular Biology and Evolution* **21**:188-199.

Swofford, D. L. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Taberlet, P., L. Gielly, G. Patou, and J. Bouvet. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* **17**:1105-1109.

Takezaki, N., A. Rzhetsky, and M. Nei. 1995. Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution* **12**:823-833.

Taylor, H. C. 1978. Capensis. Pp. 171-229 *in* M. J. A. Werger, ed. Biogeography and ecology of southern Africa. Junk, The Hague.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**:4876-4882.

Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**:689-702.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**:1647-1657.

Vinnersten, A., and K. Bremer. 2001. Age and biogeography of major clades in Liliales. *American Journal of Botany* **88**:1695-1703.

Wikström, N., V. Savolainen, and M. W. Chase. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society of London Series B-Biological Sciences* **268**:2211-2220.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *CABIOS or Computer Applications in the Biosciences* **13**:555-556.

Yoder, A. D., and Z. H. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution* **17**:1081-1090.

Young, N. D., and J. Healy. 2001. GapCoder: A computer program for including indels in phylogenetic analysis. Published privately

Tables

Table 1. Percentage of species sampled at each selected node for each sampling run. The second column gives the total number of species subtended by each selected node, the subsequent columns the percent of these species sampled.

Nodes	Total species	35 taxon	51 taxon	80 taxon	100 taxon	120 taxon	150 taxon	300 taxon
1	4	50	50	50	50	75	75	100
2	39	5	13	31	33	36	41	100
3	47	9	15	30	34	38	45	100
4	50	4	10	18	24	30	44	100
5	11	18	18	27	36	36	55	100
6	61	7	11	20	26	31	49	100
7	108	7	14	24	30	34	47	100
8	2	100	100	100	100	100	100	100
9	110	9	15	25	31	35	48	100
10	3	67	100	100	100	100	100	100
11	113	11	18	27	33	37	50	100
12	4	50	100	100	100	100	100	100
13	117	12	21	30	35	39	51	100
14	31	6	10	23	35	42	45	100
15	30	7	10	20	33	43	53	100
16	61	7	10	21	34	43	49	100
17	2	100	100	100	100	100	100	100
18	63	10	13	24	37	44	51	100
19	180	11	17	28	36	41	51	100
20	12	17	33	33	33	42	50	100
21	192	11	18	28	35	41	51	100
22	49	4	10	16	22	37	51	100
23	12	17	17	25	33	33	33	100
24	61	7	11	18	25	34	48	100
25	253	10	17	26	33	40	50	100
26	9	22	22	33	33	44	44	100
27	262	11	17	26	33	40	50	100
28	11	18	18	27	36	36	45	100
29	273	11	17	26	33	40	50	100
30	274	11	17	26	33	40	50	100
31	24	8	8	25	29	29	46	100

Table 2. Results of Modeltest, clock tests, and the penalized likelihood (PL) cross-validation procedure for each of the seven nested taxon samplings. Values with asterisk: attempts to determine optimal smoothing value failed; therefore, three zero or positive values were chosen (consistent with the range of values determined for the smaller datasets) to bracket the range of probable values.

	35 taxon	**51 taxon**	**80 taxon**	**100 taxon**	**120 taxon**	**150 taxon**	**300 taxon**
Modeltest results	GTR+G+I	GTR+G+I	GTR+G+I	GTR+G+I	GTR+G+I	GTR+G+I	GTR+I
LR test: clock	rejected (X^2= 106.5, df=33, *P*<0.01)	rejected (X^2=145.0, df=49, *P*<0.01)	rejected (X^2= 234.0, df=78, *P*<0.01)	rejected (X^2= 308.7, df=98, *P*<0.01)	rejected (X^2= 329.6, df=118, *P*<0.01)	rejected (X^2= 4830.3, df=148, *P*<0.01)	rejected (X^2= 7297.2, df=298, *P*<0.01)
PL smoothing value (log10)	6.5	5.0 (4.5-6.0 were equally optimal)	4.0	5.0	1.5	0.0, 3.5, 6.5*	0.0, 3.5, 6.5*

Table 3. Mean node ages in millions of years obtained for the four different molecular dating methods with the seven different sampling strategies. The PL values for 150 and 300 taxon were predicted from the 35-120 taxon samples by optimal regression (see methods). CL: assuming a global, constant clock; PL: Penalized likelihood; Bayes: Bayesian; NRS: non-parametric rate smoothing,

Nodes	Methods	35 taxon	51 taxon	80 taxon	100 taxon	120 taxon	150 taxon	300 taxon
1	CL	7.56	7.72	7.98	8.50	9.23	8.56	8.81
	PL	7.65	7.81	7.96	8.60	8.36	8.43	8.80
	Bayes	6.53	7.04	7.99	9.22	8.65	10.77	13.38
	NPRS	7.82	10.26	11.56	13.79	14.13	14.94	19.20
2	CL	10.48	10.61	10.77	10.91	11.09	10.68	11.10
	PL	10.52	10.58	10.69	10.92	10.14	10.23	10.67
	Bayes	9.47	10.04	11.52	12.01	12.19	13.26	15.90
	NPRS	10.40	13.26	15.08	17.96	18.58	20.28	24.81
3	CL	12.12	12.58	12.96	13.15	13.71	13.30	13.61
	PL	12.16	12.52	12.86	13.17	12.70	12.81	13.36
	Bayes	11.09	11.92	13.52	14.06	14.45	15.70	18.20
	NPRS	12.12	15.85	17.62	20.36	21.11	22.98	27.00
4	CL	12.18	13.45	13.34	13.64	14.07	13.76	14.14
	PL	12.40	13.47	13.44	13.56	12.87	12.98	13.54
	Bayes	10.79	11.55	12.81	13.09	13.46	14.80	16.76
	NPRS	11.97	15.49	16.89	18.15	19.31	21.56	25.53
5	CL	7.94	8.27	9.86	10.03	10.31	9.99	9.81
	PL	7.90	8.22	9.74	9.89	9.25	9.33	9.73
	Bayes	6.65	7.12	9.06	9.34	9.64	10.55	13.54
	NPRS	7.00	8.65	11.62	12.01	12.94	14.79	18.72
6	CL	12.77	13.45	13.59	13.64	14.07	13.76	14.14
	PL	12.74	13.47	13.66	13.56	12.87	12.98	13.54
	Bayes	11.47	12.35	13.33	13.72	14.13	15.51	18.61
	NPRS	12.36	15.49	17.16	18.15	19.31	21.56	25.53
7	CL	14.45	15.10	15.28	15.48	16.04	15.58	15.90
	PL	14.47	15.04	15.34	15.49	14.92	15.05	15.70
	Bayes	13.44	14.36	15.59	16.11	16.54	17.85	20.37
	NPRS	14.48	18.84	19.94	22.63	23.45	25.36	28.97
8	CL	9.00	9.18	9.33	9.36	9.71	9.43	13.45
	PL	9.16	9.38	9.49	9.53	9.31	9.39	9.80
	Bayes	8.94	9.48	10.26	10.67	10.96	11.80	13.33
	NPRS	9.76	12.73	13.51	15.56	16.34	18.06	21.83
9	CL	15.12	15.64	15.76	15.93	16.54	16.16	16.62
	PL	15.20	15.53	15.80	15.93	15.58	15.72	16.39
	Bayes	14.45	15.27	16.52	17.01	17.52	18.88	21.58
	NPRS	15.60	19.95	20.87	23.60	24.50	26.38	29.91
10	CL	8.98	9.18	9.25	9.27	9.61	9.29	9.28
	PL	9.00	9.24	9.36	9.40	9.27	9.35	9.75

	Bayes	9.21	9.79	10.66	11.02	11.37	12.29	14.04
	NPRS	9.76	12.88	13.77	15.98	16.83	18.67	22.62
11	CL	15.63	16.09	16.21	16.40	17.05	16.66	17.05
	PL	15.78	16.06	16.30	16.41	16.13	16.27	16.97
	Bayes	15.38	16.14	17.42	17.96	18.53	19.93	22.76
	NPRS	16.41	20.81	21.70	24.51	25.42	27.26	30.70
12	CL	14.20	13.61	14.06	13.90	14.60	14.13	14.05
	PL	14.20	13.86	14.00	14.08	14.38	14.51	15.13
	Bayes	14.50	14.92	16.11	16.72	17.16	18.34	20.16
	NPRS	15.62	19.54	20.43	23.20	24.15	25.95	29.35
13	CL	16.66	17.13	17.64	17.53	18.38	17.96	18.12
	PL	16.80	17.18	17.50	17.61	17.69	17.85	18.61
	Bayes	16.83	17.77	19.13	19.78	20.36	21.77	24.46
	NPRS	17.96	22.65	23.53	26.42	27.34	29.07	32.26
14	CL	10.92	11.23	12.39	13.30	14.25	13.81	12.71
	PL	10.92	11.54	12.82	13.50	14.22	14.35	14.96
	Bayes	11.66	12.60	14.19	15.43	15.99	16.86	18.53
	NPRS	12.64	15.96	17.66	21.21	22.70	24.15	24.14
15	CL	14.16	14.84	15.21	15.11	15.84	15.58	15.17
	PL	14.29	14.90	15.48	15.62	16.07	16.21	16.91
	Bayes	14.55	15.34	16.35	17.21	17.63	18.55	19.73
	NPRS	16.04	20.01	20.46	23.99	25.43	26.95	29.80
16	CL	15.71	16.36	16.77	17.17	17.90	17.56	16.83
	PL	15.81	16.51	17.05	17.31	18.07	18.23	19.01
	Bayes	16.36	17.31	18.38	19.31	19.79	20.83	22.16
	NPRS	17.85	22.29	22.78	26.15	27.38	28.88	31.58
17	CL	13.57	14.33	14.43	14.35	15.16	15.95	14.58
	PL	13.68	14.11	14.47	14.78	15.40	15.54	16.20
	Bayes	14.58	15.43	16.48	17.09	17.50	18.43	19.11
	NPRS	15.80	19.62	20.39	23.09	24.10	25.46	28.35
18	CL	18.58	19.22	19.68	19.94	21.29	20.61	19.97
	PL	18.75	19.35	19.80	20.18	20.96	21.15	22.05
	Bayes	19.35	20.35	21.70	22.49	23.06	24.25	25.67
	NPRS	21.00	25.55	26.46	29.43	30.39	31.86	34.51
19	CL	19.35	19.99	20.42	20.62	21.77	21.23	21.03
	PL	19.48	20.03	20.43	20.80	21.55	21.74	22.68
	Bayes	20.29	21.31	22.73	23.50	24.08	25.35	27.04
	NPRS	21.73	26.28	27.25	30.15	31.05	32.51	35.17
20	CL	15.59	15.83	16.05	16.17	16.95	16.30	16.16
	PL	15.52	15.83	16.02	16.31	17.05	17.20	17.94
	Bayes	17.02	17.95	19.08	19.78	20.38	21.55	23.09
	NPRS	18.10	22.14	22.88	25.66	26.65	28.26	32.03
21	CL	20.24	20.91	21.15	21.60	22.51	21.98	21.89
	PL	20.20	20.95	21.02	21.62	22.23	22.43	23.39
	Bayes	21.47	22.36	23.82	24.69	25.19	26.48	28.47
	NPRS	22.84	27.25	28.10	30.98	31.82	33.22	35.86
22	CL	12.59	14.62	14.61	15.52	16.12	15.48	15.21
	PL	12.74	14.81	14.85	15.68	15.26	15.39	16.06
	Bayes	13.63	14.96	16.10	16.92	17.55	18.32	22.19
	NPRS	14.72	17.74	18.34	22.07	22.49	24.31	27.53

23	CL	13.04	13.56	13.59	14.49	15.02	14.57	14.93
	PL	13.20	13.97	13.63	14.59	15.57	15.71	16.38
	Bayes	15.43	16.41	17.75	18.42	18.85	19.96	18.29
	NPRS	16.44	20.18	21.40	24.72	25.46	27.21	30.74
24	CL	17.68	19.02	19.37	20.27	21.07	20.84	20.80
	PL	17.86	19.11	19.45	20.29	20.72	20.90	21.80
	Bayes	19.59	20.70	22.29	23.27	23.71	25.21	27.01
	NPRS	20.93	25.24	26.33	29.45	30.19	31.86	34.56
25	CL	20.94	21.88	22.00	22.61	23.56	23.03	23.15
	PL	20.98	21.77	21.93	22.58	23.25	23.46	24.46
	Bayes	22.50	23.54	25.03	25.97	26.49	27.78	30.07
	NPRS	23.70	28.28	29.44	31.98	32.80	34.13	36.69
26	CL	18.11	18.75	18.32	18.76	19.93	19.37	19.88
	PL	18.20	18.64	18.40	18.68	19.98	20.16	21.02
	Bayes	18.97	19.83	21.23	22.06	22.88	24.00	26.16
	NPRS	20.10	24.37	25.39	28.24	29.23	30.66	33.87
27	CL	21.25	22.15	22.23	22.90	23.79	23.20	23.36
	PL	21.28	22.03	22.14	22.78	23.47	23.68	24.70
	Bayes	23.14	24.17	25.68	26.61	27.14	28.44	30.97
	NPRS	23.99	28.55	29.44	32.22	33.04	34.35	36.69
28	CL	22.89	23.32	21.86	22.26	22.91	22.36	22.50
	PL	22.78	22.79	21.58	22.24	23.75	23.96	24.99
	Bayes	26.07	26.69	27.25	27.95	28.17	29.09	30.33
	NPRS	26.91	30.61	30.55	32.65	33.37	34.56	37.31
29	CL	27.10	27.77	27.71	28.39	29.10	28.48	28.24
	PL	26.90	27.28	27.36	28.01	29.15	29.41	30.67
	Bayes	29.99	30.74	31.89	32.86	33.10	34.06	35.49
	NPRS	30.37	34.07	34.83	36.93	37.54	38.51	40.31
30	CL	29.35	29.72	29.80	30.26	31.07	30.47	30.12
	PL	29.44	29.67	29.74	30.23	31.31	31.59	32.95
	Bayes	32.45	33.08	34.18	35.04	35.33	36.21	38.36
	NPRS	32.96	36.16	36.92	38.68	39.24	40.06	41.66

Figures

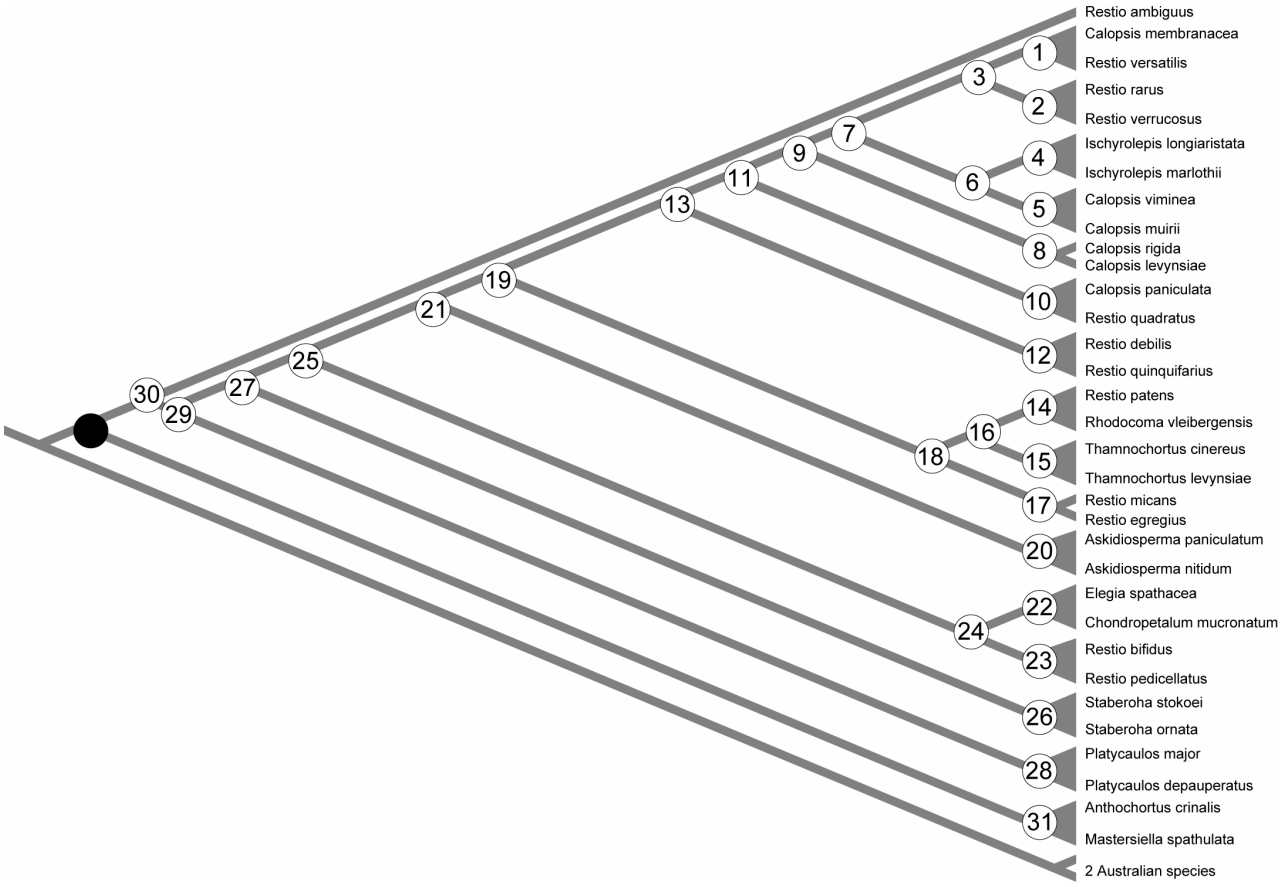


Figure 1. Cladogram used in this study. Species listed are those used in the 35 taxon analysis. The black circle is the calibration node, the numbered open circles are the test nodes.

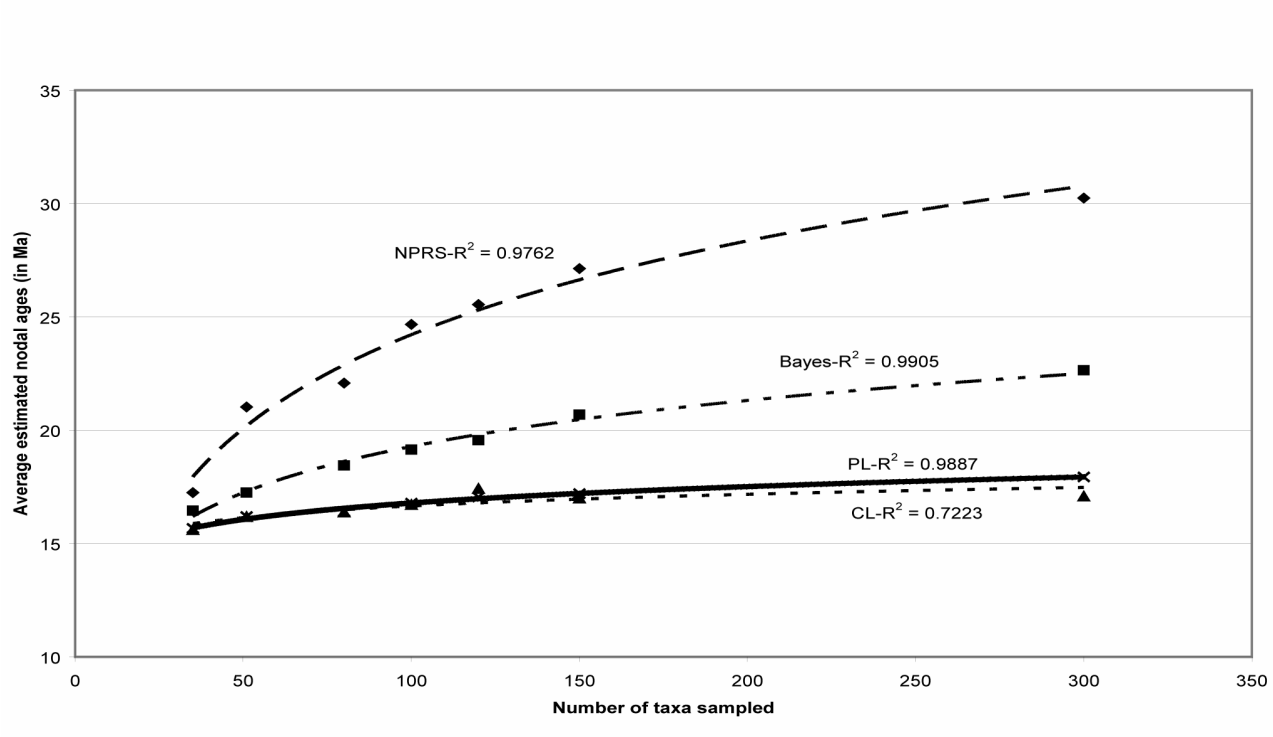


Figure 2. A comparison of the actual ages returned by the four different molecular dating methods for the seven nested taxon samples. The x axis gives the number of taxa sampled, the y axis the average age of the 30 test nodes. Diamonds: NPRS, squares: Bayesian; crosses: PL and triangles: CL.

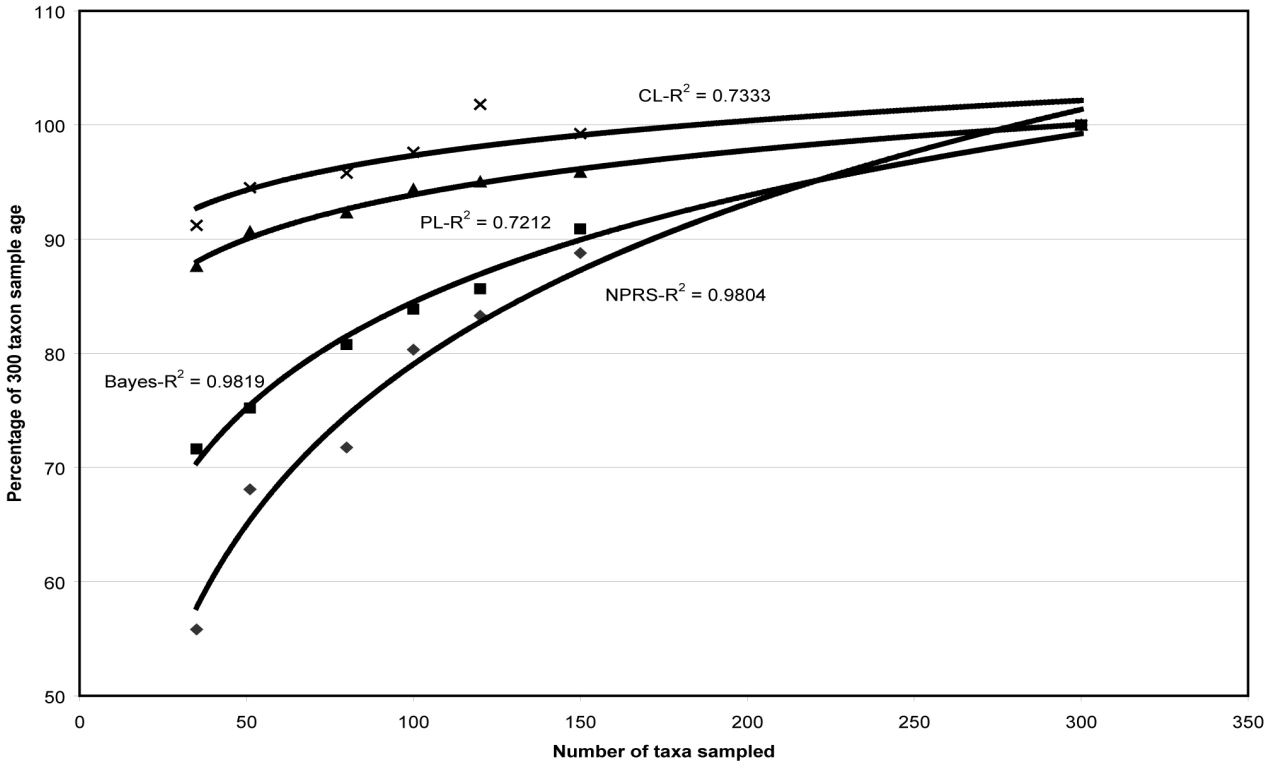


Figure 3. Effect of the species sampling density on the average test node age estimated, for the four different molecular dating methods. The x axis is the taxon sample size. The y axis indicates, for each method, the average proportion of the 300 taxon sample age obtained for each species sample size.

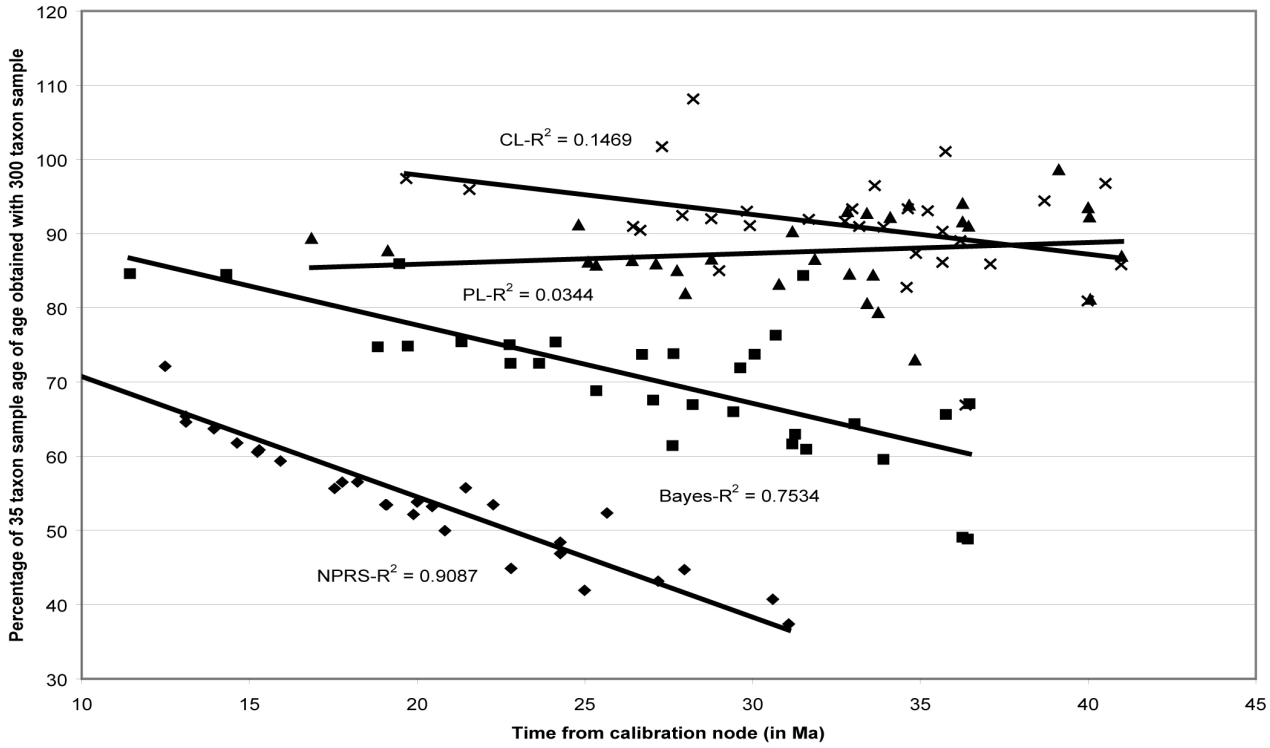


Figure 4. Proportion of node age underestimation relative to distance from the calibration point. The x axis indicates the time between the test node and the calibration node, calculated with each method based on the 300 taxon sample. The y axis represents the proportion of the age obtained with the 35 taxon sample of the age obtained with the 300 taxon sample, for each method of analysis. For both Bayes and NPRS the slope deviates significantly (at $p < 0.01$) from 0, and the r^2 is significant at the same p-value.

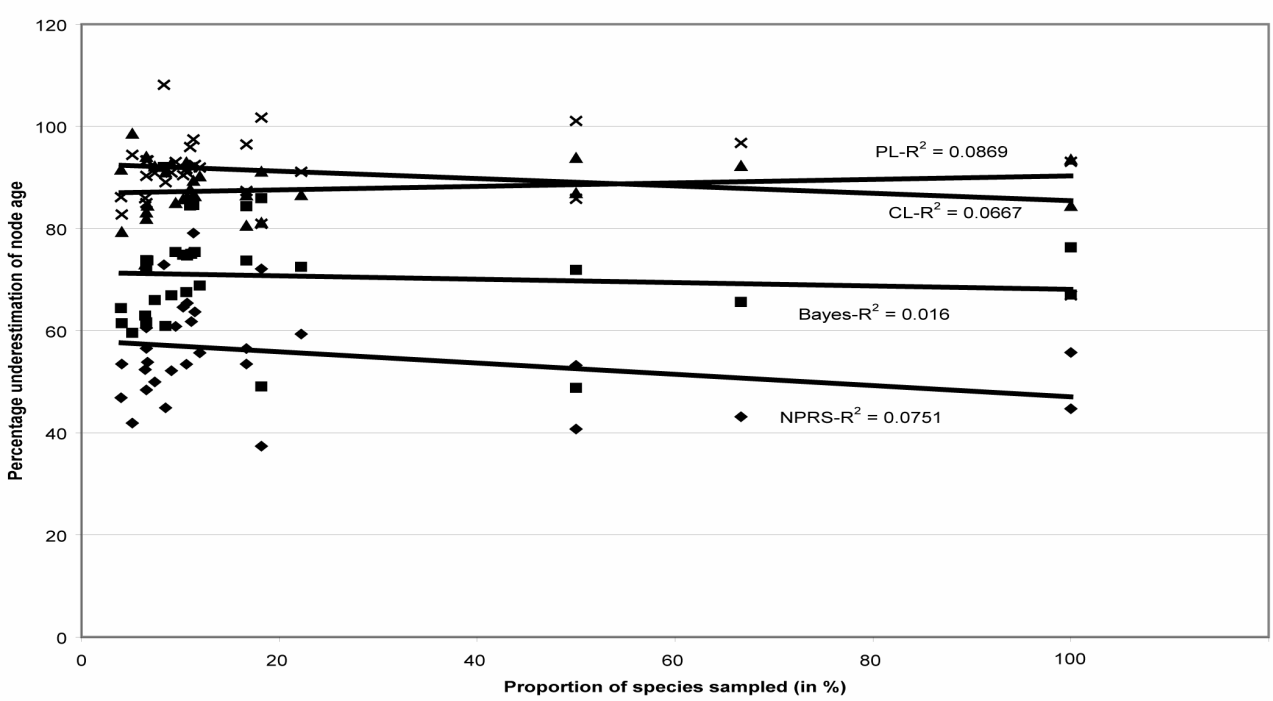


Figure 5. Differential effects of sampling on each test node individually in the 35 taxon analysis, for the four different molecular dating methods used. The x axis indicates the proportion of the species sampled above each node. The y axis represents the percentage of age undersampling for each node.

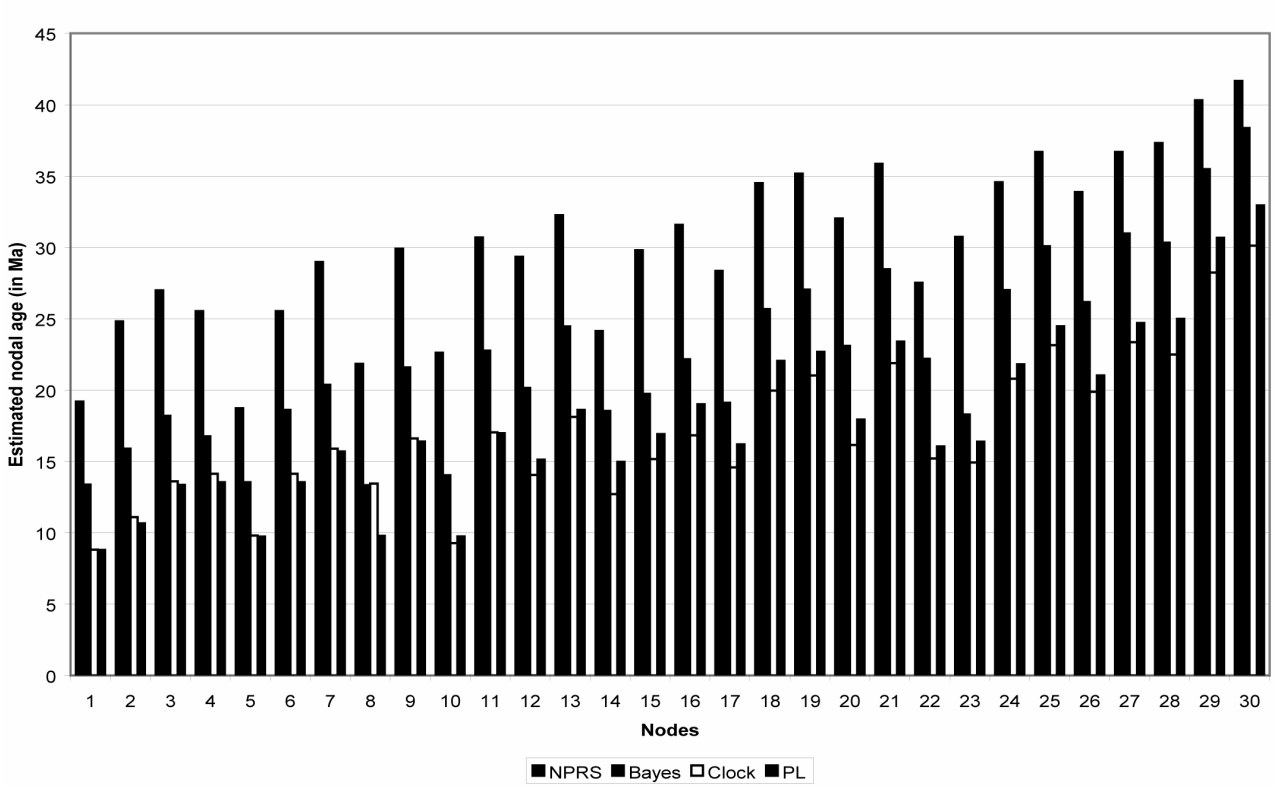


Figure 6. Comparison of the average nodal ages estimated by the four different molecular dating methods for the 300 taxon sample. The nodes are numbered as in Figure 1.

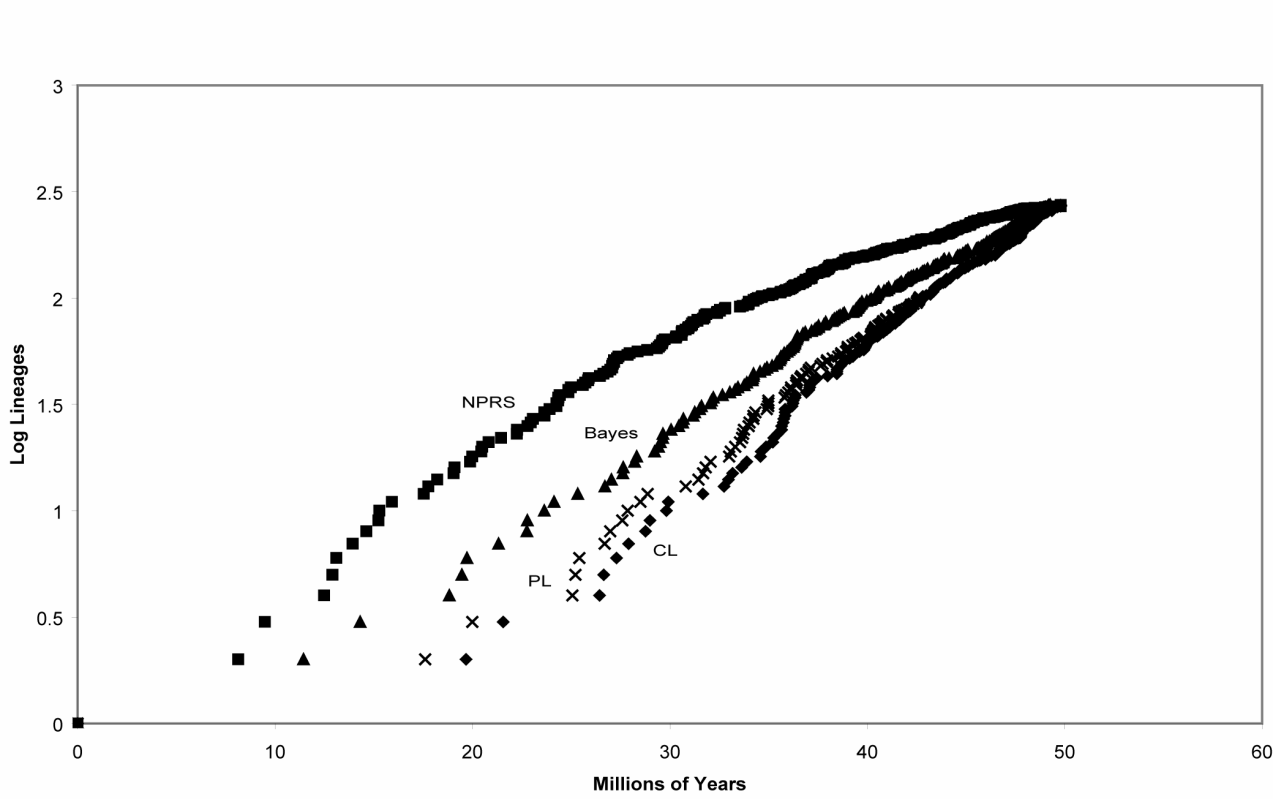


Figure 7. Lineage through time plots for 300 taxa, using four different molecular dating methods. Squares: NPRS, triangles: Bayesian, crosses: PL, and diamonds: CL. Note that the PL values for the 300 taxon sample were estimated.

General summary

The six chapters that compose this dissertation are all related to various aspects of molecular dating, ranging from more methodological and experimental work to the application of different methods in the context of biological and evolutionary questions and hypotheses.

The first chapter represents an up-to-date review on the most common molecular dating methods, which provides practical information and useful comparison tables intended for researchers looking for their method of choice and its advantages and drawbacks. In addition, the review provides links to the most recent and important literature on molecular dating.

In the second chapter, we used evidence derived from molecular dating to test the out-of-India hypothesis for Crypteroniaceae, a small group of Southeast Asian tropical rainforest trees. According to this biogeographic scenario, Crypteroniaceae were among other Gondwanan taxa that have been carried from Gondwana to Asia by rafting on the Indian plate as it moved along the African continent, from where they dispersed "out-of-India" into South and Southeast Asia, after India collided with the Asian continent in the Early Tertiary. We tested the validity of this hypothesis by inferring the age of Crypteroniaceae based on Maximum Likelihood analyses of three chloroplast DNA regions (*rbc*L, *nhd*F, and *rpl*16 intron) and using both clock-based (Langley-Fitch) and clock-independent age estimates (Non Parametric Rate Smoothing and Penalized Likelihood).Our dating results indicated an ancient Gondwanan origin of Crypteroniaceae in the Early to Middle Cretaceous, followed by diversification on the Indian plate in the Early Tertiary and subsequent dispersal to Southeast Asia. These findings were congruent with recent molecular, paleontological, and biogeographic results in vertebrates.

The third chapter represents a reply to Robert G. Moyle (2004), who reanalyzed data published in an earlier paper on the out-of-India hypothesis for Crypteroniaceae (Conti *et al.*, 2002). By using a different taxon sampling and another calibration point than Conti *et al.* (2002), Moyle concluded that the out-of-India hypothesis for Crypteroniaceae had to be rejected. In our reply, we defended the analysis by Conti *et al.* (2002), highlighted some weaknesses in Moyle's analytical procedure, and offered in the same time some general reflections on the controversial issue of calibration in molecular dating analyses.

The fourth chapter is about the relatively small group of trees and shrubs which belong to the South and East African Penaeaceae, Oliniaceae, and Rhynchocalycaceae, all closely related to the Crypteroniaceae treated in chapters two and three (order Myrtales). Chronograms inferred in previous studies revealed long stem branches for both Penaeaceae and Oliniaceae, indicating a long time period between the origin and diversification of these lineages. The chronograms also

suggested that this long phase was followed by a relatively short episode with rapid diversification, especially within Penaeaceae. In order to reconstruct the tempo, mode, and possible reasons of diversification of these plant groups, we performed phylogenetic and molecular dating analyses based on eight chloroplast and three nuclear gene sequences, followed by statistical evaluation of lineages through time plots and Maximum Likelihood reconstructions of the ancestral character states for leaf size and vegetation type. We showed that the crown radiation of Penaeaceae started about 17 million years ago, following the initiation of a climatic trend towards the modern seasonally cold and arid conditions in the Cape Floristic Region. Our ancestral character state reconstructions indicate general shifts from tropical to fynbos vegetation types, and from macro- to leptophyllous leaves, corroborating the hypothesis that modern species, adapted to low nutrient soils and summer aridity, replaced an ancestral tropical flora during the Late Miocene aridification.

In the fifth chapter, we addressed one of the crucial challenges in molecular dating: the problem of fossil calibration, or more precisely: the uncertainty that often surrounds the assignment of fossils to specific nodes in a phylogeny. We used Bayesian molecular dating and the fossil cross-validation procedure of Near and Sanderson (2004) in a novel way to assess uncertainty in fossil nodal assignment. More specifically, by using an expanded Myrtales data set and six fossils with alternative assignments, we identified the three most congruent calibration sets by comparing the consistencies of 72 multiple calibration sets available for our study. The three selected calibration sets were characterized by lower standard deviations associated with their estimated divergence times.

Finally, in the sixth chapter, we tested the sensitivity of three commonly used molecular dating methods to taxon sampling (Non Parametric Rate Smoothing, Penalized Likelihood, and Bayesian dating with multidivtime). By using a nearly complete sample of the *Restio*-clade of the African grass-like Restionaceae (300 species, including 26 outgroup species), we formed nested subsets of 35, 51, 80, 120, and 150 species and performed molecular dating analyses based on the full dataset and all the subsets. We then compared the impact of the different dataset sizes and dating methods on the age estimates. Our dating experiments showed that all methods were sensitive to undersampling, depending on the amount of rate smoothing used in the dating analyses. In addition, we observed that the undersampling effect was positively related to the distance from the calibration node. Finally, we observed that both Penalized Likelihood and Bayesian dating methods are less sensitive to undersampling than Non Parametric Rate Smoothing.

Zusammenfassung

Die sechs Kapitel der vorliegenden Dissertation befassen sich alle mit verschiedenen Aspekten von molekularen Datierungsmethoden, englisch *molecular dating methods* genannt. Dabei reicht die Spannweite der behandelten Themen von experimentellen Untersuchungen der Methoden selbst bis hin zur praktischen Anwendung der molekularen Altersbestimmung für die Aufklärung von biologischen und evolutionsgeschichtlichen Fragen.

Das erste Kapitel stellt eine aktuelle Beschreibung der meisten heute gebräuchlichen molekularen Datierungsmethoden dar und bietet dem interessierten Leser praktische Informationen und Vergleichstabellen. So werden zum Beispiel die spezifischen Vor- und Nachteile der Methoden besprochen und Hinweise auf die aktuellste relevante Literatur zum Thema gegeben.

Im zweiten Kapitel werden molekulare Datierungsmethoden benützt, um die Verbreitungsgeschichte der Crypteroniaceen, einer kleinen Gruppe von süd- und südostasiatischen Regenwaldbäumen, zu untersuchen. Dabei wurde die sogenannte *Out-of-India*-Hypothese getestet, welche postuliert, dass die Crypteroniaceen sich in der späten Kreidezeit von Afrika (das damals Teil des Superkontinents Gondwana war) auf die indischen Kontinentalplatte ausgebreitet haben und mit dieser durch Kontinentaldrift nach Asien gelangt sind. Gemäss dieser Hypothese haben sich die Pflanzen dann später, im frühen Tertiär, als die indische Kontinentalplatte auf Asien stiess, *Out-of-India* nach Süd- und Südostasien ausgebreitet. Wir testeten dieses Szenario, indem wir das phylogenetische Alter der Crypteroniaceen bestimmten. Dazu rekonstruierten wir zuerst den Stammbaum der Crypteroniaceen mittels *Maximum Likelihood*-Analysen von drei DNA-Abschnitten des Chloroplasten-Genoms (*rbc*L, *nhd*F, und *rpl*16 intron) und benützten anschliessend drei verschiedene molekulare Datierungsmethoden (*Langley-Fitch, Non Parametric Rate Smoothing* und *Penalized Likelihood*) zur Altersbestimmung. Unsere DNA-basierten Datierungsresultate zeigten, dass der Ursprung der Crypteroniaceen tatsächlich auf die frühe bis mittlere Kreidezeit zurückgeht und dass sich diese Arten wie im Szenario beschrieben über die driftende indische Kontinentalplatte nach Asien verbreitet haben könnten. Dieser Befund wird übrigens durch neuere molekulare und paläontologische Erkenntnisse bei verschiedenen Wirbeltiergruppen gestützt.

Das dritte Kapitel stellt die Antwort auf einen kritischen Artikel von Robert G. Moyle (2004) dar, der Altersbestimmungen aus einer früheren Arbeit über die *Out-of-India*-Hypothese von Conti *et al.* (2002) wiederholte und diese anzweifelte. Durch Verwendung eines anderen Kalibrierungspunktes in seiner molekularen Datierung kam Moyle zum Schluss, dass sich die

Crypteroniaceen nicht durch Kontinentaldrift, sondern rein zufällig durch *long distance dispersal* nach Asien ausgebreitet hätten. In unserer Antwort, welche gleichzeitig mit Moyles Kritik im *International Journal of Organic Evolution* erschien, verteidigten wir die Analysen von Conti *et al.* (2002), wiesen auf einige gravierende Schwächen in Moyles Analysen hin und diskutierten das kontroverse Gebiet der Kalibrierung bei der molekularen Altersbestimmung.

Im vierten Kapitel untersuchten wir eine kleine Gruppe von nah verwandten süd- und ostafrikanischen Bäumen und Sträuchern, die Penaeaceen, Oliniaceen und Rhynchocalycaceen. Diese Pflanzen sind alle eng verwandt mit den asiatischen Crypteroniaceen (siehe Kapitel zwei und drei) und gehören zur Ordnung Myrtales. Datierte Stammbäume, sogenannte Chronogramme, zeigten in früheren Studien, dass die Penaeaceen und Oliniaceen während einer langen Zeitperiode von über 40 Millionen Jahren Dauer keine neuen Arten gebildet haben oder diese einer hohen Sterberate unterworfen waren. Erst vor etwa 20 Millionen Jahren nahm die Artenzahl dann plötzlich massiv zu. Um dieses erstaunliche Diversifikationsmuster zeitlich genauer zu untersuchen und die möglichen Ursachen dafür aufzuklären, führten wir phylogenetische Untersuchungen und molekulare Altersbestimmungen durch, die auf acht Genen aus dem Chloroplasten-Genom und drei nukleären Genen beruhten. Anschliessend stellten wir *lineages through time plots* auf und rekonstruierten mit *Maximum Likelihood*-Verfahren die mutmassliche, ursprüngliche Blattgrösse und den ursprünglich bevorzugten Vegetationstyp der Vorfahren der untersuchten Pflanzen. Wir konnten zeigen, dass die Phase der intensiven Artbildung vor etwa 17 Millionen Jahren begann und vermutlich durch eine Klimaänderung in Südafrika ausgelöst wurde, die der Kap-Region ein saisonales, mediterranes Klima mit niederschlagsreichen Wintern und heissen, trockenen Sommern brachte. Unsere Rekonstruktionen ergaben ferner, dass die Pflanzen im Verlaufe ihrer evolutiven Entwicklung die Blätter verkleinert haben und sich von einem ursprünglich tropischen Habitat an die heute vorherrschende mediterrane Fynbos-Vegetation angepasst haben. Diese Merkmalsänderungen sind als Anpassungen an die sommerliche Trockenheit und die nährstoffarmen Böden zu verstehen, die sich im späten Miozän in Südafrika entwickelten.

Im fünften Kapitel geht es um das oben bereits erwähnte und umstrittene Thema der Kalibrierung der „molekularen Uhr“ bei der Durchführung der Altersbestimmung. Da die Datierungsresultate im Wesentlichen von der Kalibrierung abhängig sind, handelt es sich um ein äusserst delikates Thema. Im Detail geht es um die Schwierigkeit, die Fossilien, die man normalerweise zum Kalibrieren verwendet und deren Alter man ungefähr kennt, präzise den zugehörigen Knotenpunkten eines Stammbaumes zuzuordnen. Wir gingen dieses Problem an, indem wir die *Bayesian dating*-Methode zusammen mit der von Near und Sanderson (2004) erstmals beschriebenen *fossil cross-validation procedure* benützten. Es gelang uns, aus 72

verschiedenen Möglichkeiten, sechs verschiedene Fossilien einem Stammbaum zuzuordnen, jene drei Kombinationen zu ermitteln, welche die konsistentesten Datierungsresultate ergaben. Als Stammbaum benützten wir eine eigens rekonstruierte Phylogenie von 74 Blütenpflanzenarten der Ordnung Myrtales. Als wir die drei ausgewählten Kombinationen zur Altersbestimmung einsetzten, stellten wir fest, dass die Datierungsresultate kleinere Standardabweichungen aufwiesen als die Ergebnisse anderer, weniger konsistenter Kombinationen von Kalibrierungspunkten.

Im sechsten Kapitel schliesslich überprüften wir, wie empfindlich drei heute verbreitet eingesetzte molekulare Datierungsmethoden (*Non Parametric Rate Smoothing*, *Penalized Likelihood* und *Bayesian dating*) auf unterschiedliches *taxon sampling* reagieren. Dazu benützten wir Datensätze verschiedener Grösse, welche DNA-Sequenzen von jeweils 35, 51, 80, 120, 150 oder 300 Restionaceen-Arten enthielten. Restionaceen sind grasähnliche Blütenpflanzen, welche hauptsächlich in Südafrika und Australien verbreitet sind. Durch Anwendung der verschiedenen Datierungsmethoden verglichen wir den Einfluss des *taxon samplings* und der verwendeten Methoden auf die Datierungsresultate. Unsere Experimente zeigten, dass alle Methoden mehr oder weniger stark auf unterschiedliches *taxon sampling* reagieren, abhängig von der Stärke des bei der Datierung eingesetzten *rate smoothings*. Der beobachtete *undersampling* Effekt war ausserdem positiv korreliert mit der Distanz zwischen den datierten Knoten im Stammbaum und dem Kalibrierungspunkt. Abschliessend stellten wir fest, dass sowohl die *Penalized Likelihood*- als auch die *Bayesian dating*-Methode weniger sensibel auf *undersampling* reagieren als die *Non Parametric Rate Smoothing*-Methode.

Curriculum vitae

Personalien

Name und Vorname:	Rutschmann, Frank Kaspar
Geboren am:	14. Januar 1970 in Bülach ZH, Schweiz
Heimatort:	Zürich ZH, Schweiz

Ausbildung

Mittelschule:	Kantonale Maturitätsschule für Erwachsene, Zürich. Abschluss 1997, Typus B (Latein).
Hochschulstudium:	Universität Zürich, Grundstudium Biologie, 4 Semester, 1997 bis 1999. ETH Zürich, Fachstudium Allgemeine Biologie, 4 Semester, 1999 bis 2001.
Diplomfach:	Allgemeine Biologie (ETH Zürich), mit den Schwerpunkten Pflanzenphysiologie, Biochemie, Molekulare Genetik, Zellbiologie und Mikrobiologie.
Diplomarbeit:	„Charakterisierung einer Calcium-abhängigen Proteinkinase *Le*CPK1 aus *Lycopersicon esculentum*“. Institut für Pflanzenwissenschaften der ETH Zürich, Prof. Nikolaus Amrhein. Herbst 2001.
Doktorat:	Universität Zürich, Institut für Systematische Botanik, Prof. Elena Conti. Assistenz (50%) von September 2002 bis Juli 2006.